Traitement automatique de vidéos en LSF Modélisation et exploitation des contraintes phonologiques du mouvement

François Lefebvre-Albaret

THESE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Informatique*

Présentée et soutenue par *François Lefebvre-Albaret*
Le 7 octobre 2010

Titre :
Traitement automatique de vidéos en LSF
Modélisation et exploitation des contraintes phonologiques du mouvement

---

JURY
Frédéric Lerasle (Président)
Patrice Dalle (Directeur)
Justus Piater (Rapporteur)
Sylvie Gibet (Rapporteur)
Dominique Boutet (Examinateur)
Pascal Jobez (Invité)

Ecole doctorale : *EDMITT*
Unité de recherche : *IRIT*
Directeur de Thèse : *Patrice Dalle*

Cette thèse a été cofinancée par la société Websourd et par la région Midi-Pyrénées.

Certaines illustrations de signes sont tirées de dictionnaires d'IVT et Monica Companys. Nous remercions ces deux éditeurs de nous avoir autorisé à les utiliser.

RÉSUMÉ

Dans le domaine du Traitement Automatique des Langues Naturelles, l'exploitation d'énoncés en langues des signes occupe une place à part. En raison des spécificités propres à la Langue des Signes Française (LSF) comme la simultanéité de plusieurs paramètres, le fort rôle de l'expression du visage, le recours massif à des unités gestuelles iconiques et l'utilisation de l'espace pour structurer l'énoncé, de nouvelles méthodes de traitement doivent être adaptées à cette langue.

Nous exposons d'abord une méthode de suivi basée sur un filtre particulaire, permettant de déterminer à tout moment la position de la tête, des coudes, du buste et des mains d'un signeur dans une vidéo mono-vue. Cette méthode a été adaptée à la LSF pour la rendre plus robuste aux occultations, aux sorties de cadre et aux inversions des mains du signeur.

Ensuite, l'analyse de données issues de capture de mouvement nous permet d'aboutir à une catégorisation de différents mouvements fréquemment utilisés dans la production de signes. Nous en proposons un modèle paramétrique que nous utilisons dans le cadre de la recherche de signes dans une vidéo, à partir d'un exemple vidéo de signe.

Ces modèles de mouvement sont enfin réutilisés dans des applications permettant d'assister un utilisateur dans la création d'images de signe et la segmentation d'une vidéo en signes.

Mots clés: phonologie, segmentation, video, patron de mouvement, LSF

ABSTRACT

There are a lot of differences between sign languages and vocal languages. Among them, we can underline the simultaneity of several parameters, the important role of the face expression, the recurrent use of iconic gestures and the use of signing space to structure utterances. As a consequence, new methods have to be developed and adapted to those languages.

At first, we detail a method based on a particle filter to estimate at any time, the position of the signer's head, hands, elbows and shoulders in a monoview video. This method has been adapted to the French Sign Language in order to make it more robust to occlusion, inversion of the signer's hands or disappearance of hands from the video frame.

Then, we propose a classification of the motion patterns that are frequently involved in the sign production, thanks to the analysis of motion capture data. The parametric models associated to each sign pattern are used in the frame of automatic sign retrieval in a video from a filmed sign example.

We finally include those models in two applications. The first one helps an user in creating sign pictures. The second one is dedicated to computer aided sign segmentation.

Keywords: French Sign Language, phonology, segmentation, video, FSL, motion pattern

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES SIGLES

ACP	Analyse en Composantes Principales
ASL	Langue des Signes Américaine
BSL	Langue des Signes Britannique
DGS	Langue des Signes Allemande
DTW	Dynamic Time Warping
LPC	Langage Parlé Complété
LSF	Langue des Signes Française
PRESTO	Pôle de Recherche Signe TOlosan
SESCA	Systèmes pour l'Enseignement de la LS et la Communication par Avatar

Il y a un peu plus de six ans . . .

– Mais que penses-tu que tu vas trouver ? Pourquoi faire de la recherche sur la Langue des Signes ?
Assis avec moi au bout de la table, Mr Markarian venait de me poser une question pour le moins inattendue. Nous avions pris le temps pour discuter tandis que les autres musique-études fêtaient joyeusement les retrouvailles.
L'homme m'avait écouté, alors que toutes les personnes avec qui j'avais parlé de mon projet de recherche autour de la LSF n'y avaient même pas prêté attention.
– J'espère vraiment que tu pourras la faire, cette recherche, m'avait-il dit. Je suis certain que tu découvriras des choses fascinantes !

C'est sur ces quelques mots que s'achevait une discussion qui m'a permis pour la première fois d'évoquer sérieusement mon rêve de marier dans mon travail, Langue des Signes et informatique.

Six ans plus tard, ce projet est devenu concret et bien réel.
– Vous aviez raison ! Ce thème est bien fascinant. La recherche m'a transportée sur des terrains encore inexplorés qui n'ont pas fini de livrer leurs secrets.

Mr Markarian, je vous dédicace cet ouvrage et vous remercie encore pour cet échange.

REMERCIEMENTS

Je n'aurai pas pu faire cette thèse sans toutes ces personnes qui m'ont fait partager leur passion pour la Langue des Signes Française. Je pense naturellement à Mme Monin, à ma marraine INSA Violaine et à tous les enseignants de Langue des Signes que j'ai eu la chance de côtoyer à Visuel Langue des Signes et à IVT. Merci à vous, ambassadeurs de la LSF qui m'avez fait découvrir une nouvelle langue toute en image.
Je voudrais également remercier tous ceux qui m'ont porté dans ma décision de choisir de m'engager dans une thèse à l'IRIT : Mr Bosc, mes amis de musique-étude, mes cousins.
Mes remerciements vont aussi aux personnes avec lesquelles j'ai travaillé ces trois années. Je n'oublierai pas ces instants de complicité avec les autres doctorants de TCI et de Vortex, ces cafés signes autour d'un "chat dans le sac" qui pouvaient vous faire oublier une dure semaine de labeur.
Un merci tout particulier à Juliette de m'avoir accueilli et introduit dans le monde des sourds de Toulouse. C'est grâce à d'ardents défenseurs comme elle que la LSF ne perd pas son âme dans les ordinateurs des chercheurs.
Ce séjour à l'IRIT aura été l'occasion de rencontrer des personnes de tous les horizons avec lesquelles j'ai pu avancer sur la route du traitement automatique des Langues des Signes. Merci à tous les enseignants d'IRIS, du SESSD et des classes bilingue d'avoir activement collaboré au développement des applications.
Merci aussi à tous les collègues chercheurs dans nos laboratoires partenaires pour ces échanges fructueux.
Merci à tous mes collègues de Websourd et plus particulièrement à Pascal, Rémi, Jérémie et Sophie qui m'ont aidé dans ma réflexion sur les signeurs virtuels et leur application.
Merci à Dominique et Michael de m'avoir accueilli et apporté leurs précieux conseils lors de la rédaction de ma thèse.
Merci enfin à Patrice, mon encadrant qui a pris le temps lorsque c'était nécessaire pour m'aider à prendre du recul et me guider dans ce labyrinthe qu'est la recherche.
Une thèse, c'est une aventure avec des hauts et des bas. Merci à tous ceux qui m'ont soutenu au cours de ces trois ans : mes amis de Toulouse et d'ailleurs, les compagnons de Bonne Nouvelle Quart Monde, les Zickets et ma famille.

CHAPITRE 1

INTRODUCTION

1.1 Contexte du travail

L'histoire de la Langue des Signes Française (LSF[1]) a connu récemment un tournant décisif grâce à l'adoption du la loi de 2005, pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées. Le texte confère à la LSF le statut de langue qui peut être enseignée ou être utilisée dans le cadre de l'enseignement. Par ailleurs, les lieux publics se doivent également de rendre leurs informations accessibles au public sourd, et la traduction en LSF de ces informations est une des options possibles pour l'accessibilité.
La création de produits permettant de répondre à ces besoins (télé-interprétation, outils pédagogiques, signeurs virtuels ...) met en lumière des besoins plus fondamentaux d'outils informatiques pour traiter ou générer des vidéos en LSF. Le fait d'enseigner la LSF rend aussi nécessaire une meilleure formalisation linguistique des spécificités des langues signées.
Le développement de nouvelles méthodes et de nouveaux modèles dédiés à la LSF implique nécessairement une collaboration entre plusieurs disciplines. Pour cette raison, les échanges entre linguistes, traiteurs d'images, spécialistes dans la génération de gestes communicatifs, spécialistes du mouvement et industriels au contact de la population sourde signante se font de plus en plus fréquents. Nous avons eu la chance de participer aux projets suivants qui s'inscrivent dans cette perspective et ont permis de développer la plupart du contenu de cette thèse :

- Le projet SESCA (Système pour l'Enseignement de la langue des Signes et la Communication par Avatar) dans lequel ont participé l'UTM, Websourd, IRIS, Interprétis et l'IRIT a permis de développer des outils pédagogiques (AVV, SLAnnotation) dédiés à l'enseignement et à l'analyse de la LSF. Ils ont également permis de développer plusieurs pistes pour articuler les modèles d'analyse et de synthèse.

- Le projet SignCom (Communication en langue des signes française entre agents réels et virtuels) où participent le VALORIA, l'IRISA de Rennes, l'Université de Rennes 2, l'IRIT, l'IRIS, Websourd et Polymorph, a été l'occasion d'acquérir un corpus conséquent de signes par capture de mouvement optique. Ces données de capture de mouvement nous ont permis de valider les modèles de mouvement que nous exposons dans cette thèse.

- Le projet UVED (Université Virtuelle pour l'Environnement Durable) où participent l'IRIS, l'IRIT et Interprétis, nous a permis d'approfondir la notion d'outil de traitement de signe et de pousser plus

[1]Nous utilisons l'acronyme LSF lorsque nos propos ne concernent que la Langue des Signes Française ou lorsqu'une étude serait nécessaire pour vérifier s'ils s'appliquent aussi aux autres langues des signes. Nous utilisons la notation LS pour parler des langues des signes en général.

loin la réflexion sur la place de la représentation graphique des signes dans les documents bilingues.

Tous ces projet ainsi que les différentes collaborations que nous avons eues avec les équipes du LIMSI, avec les linguistes de Paris VIII et avec l'Université du Québec à Montréal, ont permis de développer plusieurs approches complémentaires pour arriver aux méthodes de traitement de vidéo que nous présentons dans ce mémoire.

1.2 Objectifs de l'étude

Quelques travaux de recherche informatique ont déjà porté sur la reconnaissance de petits nombre de signes isolés ou en contexte. Il s'agissait souvent de traduire des énoncés simples d'une Langue des Signes (LS) vers une Langue Vocale (LV).

Les travaux que nous présentons dans cet ouvrage ont un but un peu différent. Il s'agit d'extraire des informations, de vidéos réelles contenant des énoncés en LS qui n'ont pas été réalisés spécifiquement dans un but d'un traitement automatique ultérieur. Nous ne restreignons pas le lexique que nous sommes en mesure de traiter, mais nous renonçons par la même occasion à la reconnaissance de signe, en nous limitant à une détection et à une caractérisation du mouvement des signes.

Lorsqu'on aborde le traitement des LS sous un angle de traiteur d'image, les problèmes à surmonter sont extrêmement nombreux. Il faut souvent développer des méthodes différentes pour extraire les expressions du visage, la posture du signeur et les configurations manuelles. D'autre part, il est possible d'effectuer des traitements à différents niveaux allant de celui d'un suivi simple des différentes parties du corps lors de l'expression d'un signeur à l'interprétation d'un énoncé. Nous avons choisi de nous focaliser sur la modélisation et l'extraction du mouvement dans des vidéos en LSF en nous cantonnant à une analyse de l'énoncé au niveau lexical. Notre but est de créer des modèles de mouvement adaptés au traitement automatique de vidéos, et de les utiliser ensuite pour effectuer un traitement de plus haut niveau sur les LS permettant de mener à terme : une recherche de signe dans une vidéo basée sur le mouvement, une représentation graphique automatisée des mouvements impliqués dans les signes et une segmentation des énoncés en signes.

Ces fonctionnalités sont ensuite utilisées dans des logiciels dédiés à l'enseignement de la LSF ou à la mise en forme de documents en LS.

1.3 Présentation de la problématique

La tâche de traitement automatique du mouvement à partir d'une simple vidéo monovue est rendue complexe à cause de difficultés comme les occultations, la vitesse des mouvements à suivre, les faibles résolutions des images à traiter ou les conditions d'acquisition de vidéo. Dans un tel contexte, il est indispensable d'ajouter à tous les niveaux de traitement, des connaissances sur le comportement de l'objet à traiter. Nous savons par ailleurs, d'après les études linguistiques, qu'il existe de nombreuses règles ou tendances régissant la structure des signes utilisés dans une production d'un énoncé en LS. Ces règles phonologiques du mouvement concernent aussi bien les déplacements des mains du signeur, prises indépendament

l'une de l'autre, que les relations entre les mouvements des deux mains durant la réalisation d'un signe. Le but de notre travail est d'étudier ces différentes règles phonologiques et de les utiliser lors des différentes étapes de traitement, allant du suivi dans la vidéo à une caractérisation de plus haut niveau des unités lexicales. Nous cherchons donc à répondre à la question suivante :

Comment peut-on intégrer les règles phonologiques régissant le mouvement des signes standards de la LSF, dans un système de traitement automatique de vidéos d'énoncés en LSF ?

1.4 Structure de la thèse

Cette thèse s'adresse aussi bien aux informaticiens, aux linguistes et à toute personne désireuse d'en savoir plus sur le Traitement Automatique des Langues des Signes (TALS). Pour cette raison, nous tenons à présenter à la fois les fondements des modèles linguistiques appliqués aux langues des signes (partie 1) et les principes des méthodes de traitement d'images utilisées pour le suivi et la reconnaissance de gestes communicatifs (partie 2). Nos contributions (partie 3) portent sur l'intégration des contraintes phonologiques du mouvement à différentes étapes du traitement de la vidéo.
Notre thèse sera donc structurée en trois parties :

- La première aborde les spécificités de la Langue des Signes Française. Ceci permettra au lecteur peu habitué au vocabulaire linguistique de se familiariser avec les termes relatifs à la LSF, que nous emploierons dans le reste de cet ouvrage. Nous en profiterons pour aborder de manière plus détaillée les différents modèles utilisés par les linguistes pour décrire les signes.

- La seconde partie présente un tour d'horizon sur les différents travaux accomplis dans notre domaine de traitement automatique de la vidéo. Nous ferons d'abord le point sur les différents travaux relatifs au suivi de gestes communicationnels en nous focalisant davantage sur les méthodes monovues. Nous présenterons ensuite un état de l'art sur la caractérisation et la reconnaissance de signes de langues des signes en mettant en avant les différentes stratégies utilisées dans le domaine du TALS.

- La troisième partie présente les méthodes que nous avons mises en œuvre pour utiliser les modèles de mouvement à différents niveaux. Nous exposons un algorithme de suivi utilisant des spécificités de la LSF. Les données de suivi obtenues grâce à la capture de mouvements nous permettent ensuite d'établir des modèles de signes paramétriques capables de représenter la plupart des signes utilisés en LS. Nous montrons finalement comment ces modèles de signes peuvent être utilisés pour corriger le suivi, effectuer une recherche par le contenu dans un document en LS, réaliser des représentations graphiques de signes ou segmenter un document en signes.

Pour plus de concision et de lisibilité, nous avons fait le choix de faire figurer les différentes méthodes calculatoires en annexes. Ainsi, les lecteurs souhaitant en savoir plus sur les filtres particulaires, les images

intégrales, les modèles de Markov Cachés et les Déformations Temporelles Dynamiques pourront se reporter à la fin du document.

Première partie

La Langue des Signes Française et sa modélisation

CHAPITRE 2

LA LANGUE DES SIGNES FRANÇAISE

Notre travail de recherche porte sur la Langue des Signes Française (LSF). Cette langue s'inscrit dans une histoire dont les derniers épisodes ont eu un impact décisif sur sa diffusion et l'état actuel de la recherche sur la LSF. Nous détaillerons dans la section qui suit les étapes marquantes du développement de la LSF en mettant en évidence les nouveaux besoins qui ont émergé du fait des récentes lois qui ont promu la diffusion et l'accès à la LSF. La synthèse présentée est basée principalement sur l'ouvrage [MVGD97b] publié par IVT auquel le lecteur avide de plus de détails pourra se reporter.

2.1 Aspects historiques

2.1.1 La naissance et l'âge d'or de la LSF

Il est extrêmement difficile pour un enfant né sourd d'apprendre à manier convenablement la langue vocale qui l'environne car il n'y a pas accès par le biais de l'audition. Les Langues des Signes peuvent donc être considérées comme un mode de communication naturel utilisé par les sourds pour exploiter le canal visuo-gestuel dont ils disposent. Le fait que les modalités de cette langue soient différentes de celles des langues vocales ne remet pas en cause son statut de langue dans la mesure où elle en remplit toutes les fonctions et où elle est capable de véhiculer un message de complexité identique.

Si cette idée semble de plus en plus acceptée à notre époque, il aura fallu plusieurs siècles pour que la LSF gagne peu à peu ce statut de langue. Dans la Grèce antique où le mot "logos" désignait à la fois la notion de pensée et de parole, le sourd qui ne pouvait pas manier une langue vocale était considéré comme incapable d'exprimer sa pensée de manière intelligible. Cet héritage permet certainement d'expliquer pourquoi les gesticulations des sourds ont été considérées au moyen-âge plus comme une preuve de folie ou de manque d'intelligence que comme un moyen de communication gestuel d'une rare sophistication.

Il faudra attendre la renaissance et le siècle des lumières pour que l'idée que les Langues des Signes constituent bien des langues à part entière voit le jour. On pourra d'ailleurs lire dans les écrits de Montaigne datant du XVI^e siècle : "Nos muets discutent et content des histoires par signes. J'en ai vu de si souples et formées à cela qu'à la vérité, il ne leur manque rien à la perfection de se faire entendre". Dès lors, il parait possible d'utiliser cette langue pour éduquer les enfants sourds. Les premières tentatives ont pour but d'utiliser les signes pour apprendre aux enfants à oraliser.

La première utilisation attestée d'une langue des signes comme langue d'enseignement est l'oeuvre d'Etienne de Fay (1610-1750), un sourd d'Amiens. Il s'agit là d'un cas isolé et il faudra attendre l'intervention de l'Abbé de l'Epée (1712-1789) pour que l'éducation en langue des signes ait lieu à plus grande

échelle. Sa méthode est basée sur l'observation des signes utilisés par les enfants sourds, et la consignation de ces signes par écrits. Il ajoute également d'autres signes correspondant à des mots du français comme les articles, prépositions, verbes d'états ... Ces "signes méthodiques" ajoutés artificiellement aux langues gestuelles étaient sensés permettre un apprentissage plus aisé du français écrit. Leur utilisation fut un échec, car ils sont en opposition avec la grammaire naturelle spatiale de la LSF. La méthode de l'Abbé de l'Epée fait école en Europe et se propage rapidement à la fin du 18ème siècle.

Les successeurs de l'Abbé de l'Epée abandonnent progressivement les signes méthodiques pour une sorte de français signé, pour revenir enfin à une Langue des Signes plus naturelle. Certains sourds deviennent enseignants. On pourra citer par exemple Jean Massieu (1772-1846), Laurent Clerc (1785-1869) et Ferdinand Berthier (1803-1886).

En 1817, apparaît grâce à Roche Ambroise Auguste Bébian (1786-1834), enseignant à l'Institut National des Jeunes Sourds de Paris, le concept révolutionnaire d'éducation bilingue. La LSF est la première langue des sourds et sa maîtrise permet l'apprentissage ultérieur du français écrit. Petit à petit, en raison du manque de professeurs maîtrisant la LSF et de l'idée qui se répandait que la langue des signes favorisait le repli de la communauté sourde sur elle-même et empêchait l'apprentissage du français, les pédagogies oralistes (basées sur le français oral) se substituent aux pédagogies bilingues.

2.1.2 Le congrès de Milan et ses conséquences

En 1880, le congrès de Milan réunit 164 participants (dont deux sourds) impliqués dans l'éducation des sourds. Suite à des démonstrations savamment organisées pour démontrer l'efficacité de la méthode oraliste, la Langue des Signes Française est officiellement interdite en France.

En 1909, le psychologue Binet, écrit un article, après enquête, sur les résultats de la méthode orale pure sur le terrain français. Il conclut en soulignant l'échec total de cette pédagogie. En effet, aucune socialisation ne s'effectuait tant dans un milieu entendant que dans un milieu sourd. De plus la lecture labiale était restreinte au cercle familial [Bin09].

L'interdiction de la langue des signes a eu plusieurs conséquences dramatiques. La première d'entre elle se trouve certainement au niveau de l'éducation des sourds. Comme le faisait déjà remarquer Ferdinand Berthier (professeur sourd) en 1853 : "Si l'éducation des sourds-muets devait se résumer dans l'articulation, la lecture sur les lèvres ou même la dactylologie, on ne pourrait commencer à leur enseigner une science [...] que lorsqu'ils seraient assez avancés dans l'étude de la langue pour comprendre les explications qu'on aurait à leur donner par cette voie". L'éducation des sourds étant focalisée sur l'oralisation, le niveau de culture des enfants sourds scolarisés durant la première moitié du $XX^{ième}$ siècle reste donc assez faible. La deuxième conséquence concerne la perte d'identité culturelle de la communauté sourde dont la langue est l'un des piliers, même si la Langue des Signes Française a toujours continué à être pratiquée hors des

institutions (et en cachette dans les institutions) jusqu'à sa réintroduction officielle dans l'éducation. Une autre conséquence qui nous touche de plus près est le caractère récent de la recherche sur la LSF. Ceci explique le peu d'ouvrages sur la grammaire de cette langue et le peu d'outils linguistiques adaptés à son étude.

2.1.3 La LSF aujourd'hui

Si c'est par l'éducation que la Langue des Signes Française est née puis s'est développée, c'est aussi par elle que le problème de Langue des Signes est reposé à la fin du XX^{ième} siècle.
En 1991, la loi Fabius reconnaît aux familles "la liberté de choix entre une communication bilingue - langue des signes et français - et une communication orale" dans l'éducation des jeunes sourds. Plus récemment, les décrets du 21 décembre 2005 (n°2005-1617) sur l'aménagement des concours, la circulaire du 27 mars 2006 (n°2006-051) sur l'enseignement de la LSF à titre optionnel et le décret du 3 mai 2006 (n°2006-509) sur le parcours scolaire des jeunes sourds réaffirme l'importance de la place de la Langue des Signes Française dans l'enseignement.
Parallèlement, la loi sur l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées du 11 février 2005 reconnaît enfin la Langue des Signes Française comme une langue à part entière et stipule que "toute personne sourde bénéficie du dispositif de communication adapté de son choix", devant les juridictions administratives, civiles et pénales. Elle réaffirme le droit des familles de choisir une éducation bilingue et précise que ce droit concerne aussi le parcours scolaire. Il revient donc à l'éducation nationale de mettre en place des dispositifs bilingues ainsi qu'une pédagogie adaptée.

Actuellement, la liberté de choix de langue d'enseignement reste difficilement applicable car le nombre d'établissement bilingues (Français - LSF) reste de loin insuffisant par rapport au nombre de demandes de scolarisation des familles. L'accessibilité devient encore plus problématique dans les établissements supérieurs où les interprètes ne sont pas assez proposés aux étudiants sourds.
De même, l'accessibilité des différentes administrations aux personnes sourdes signantes peut encore être beaucoup améliorée, même si des progrès sont déjà effectués en la matière avec la formation du personnel chargé de l'accueil ou avec les dispositifs de télé-interprétation.

Il n'en demeure pas moins que le récent changement de statut de la LSF fait émerger de nombreux besoins dans plusieurs domaines :

Au niveau de l'enseignement en LSF , les enseignants manquent d'outils pédagogiques adaptés à la transmission d'un savoir en langue des signes. Il est nécessaire de chercher avec eux des solutions pour fabriquer des supports pédagogiques bilingues, commenter un document en LSF, ou corriger une production signée.

Au niveau de l'enseignement de la LSF , les besoins sont aussi colossaux. En effet, les outils consacrés aux langues vocales et écrites ne sont pas forcément transposables à cause de la nature spatiale des

Langues des Signes. Il faut donc inventer des outils plus visuels pour rendre compte de la grammaire des Langue des Signes et inventer de nouveaux outils pédagogiques (notamment pour représenter le lexique).

Au niveau de la diffusion de la LSF . Vu le nombre d'interprètes et le faible nombre de sourds signants, il n'est pas réaliste d'affecter un interprète dans chaque lieu recevant du public. Des solutions sont alors à trouver pour avoir accès à une interprétation du message à la demande. Actuellement, différentes sociétés (Websourd, Tadeo et Viable) développent dans cette optique des outils pour la télé-interprétation. Lorsque le type de message devient récurent comme dans les gares SNCF, il est possible de générer automatiquement des messages par le biais d'avatars signants.

Notre thèse s'inscrit dans cette logique de développement d'outils autour de l'enseignement de/et en Langue des Signes Française ainsi que le traitement automatique de la Langue des Signe en vue d'en effectuer une synthèse par avatar signant.

2.2 Quelques définitions

Avant de rentrer plus précisément dans la description des spécificités de la Langue des Signes Française, il est important de bien définir ce terme et de lever l'ambiguïté quant à d'autres modes de communication faisant aussi appel à la modalité gestuelle.

2.2.1 Langue des Signes Française

La langue des Signes Français sur laquelle porte notre travail est une langue à part entière reconnue comme telle par la loi française (cf. §2.1.3). Elle dispose d'un vocabulaire et d'une syntaxe autonome de la langue française. Contrairement à des idées erronées, la LS n'est pas internationale. On compte souvent une, voire plusieurs langues des signes par pays. L'unité de la LSF à l'échelle de la France porte sur la syntaxe et une portion importante de signes standards unifiés même s'il n'est pas rare de découvrir certaines variantes régionales. De même, les différences entre les langues des signes portent essentiellement sur le lexique et il est possible de trouver un grand nombre de similarités dans leur syntaxe spatiale et temporelle. Actuellement, on peut estimer à entre 300 000 et 500 000 le nombre de personnes sourdes pratiquant la LSF, même si ce chiffre recouvre de grandes disparité dans la maîtrise de la langue.

2.2.2 Dactylologie ou alphabet manuel

La dactylologie est l'alphabet manuel utilisé dans le cadre de la LSF. Il sert à épeler les noms n'ayant pas ou pas encore d'équivalents signés de la Langue des Signes Française (certains nom propres par exemple). On trouve les configurations manuelles de la dactylologie dans de nombreux signes du lexique de la LSF. Naturellement, une langue des signes ne consiste pas à épeler les mots du français un à un. L'usage de l'alphabet manuel est donc très marginal dans la production d'un énoncé signé.

2.2.3 Langage Parlé Complété

Le Langage Parlé Complété (LPC) est un codage phonétique du français employé en complément de la lecture labiale. Les configurations manuelles sont exécutées avec une main à proximité du visage pour rendre distincts des phonèmes ayant la même empreinte labiale (ex : ma/ba/pa). Le LPC ne constitue pas une langue dans le sens où il n'est en rien indépendant du français. Le travail de recherche que nous présentons ne concerne en aucun cas le LPC.

2.2.4 Français signé

Le français signé utilise les signes de la LSF pour effectuer une traduction plus ou moins linéaire du français. Si cette définition reste imprécise, c'est que cette manière de communiquer effectue un compromis entre la syntaxe du français et celle de la Langue des Signes. Un français signé pourra donc aller d'une traduction signée mot-à-mot d'un texte français à une langue des signes fortement influencée par le français dans sa syntaxe et son manque d'iconicité[1].

2.3 Une organisation différente des langues vocales

Avant d'entrer dans le détail des techniques de traitement automatique des vidéos en langue des signes, il est important d'en souligner les spécificités qui font de la LSF une langue unique. Il ne s'agira donc pas d'une présentation détaillée des structures grammaticales de la LSF mais plutôt d'une précision de notre cadre de travail.

La langue des Signes Française partage avec les langues orales[2] de nombreux points communs dans les fonctions qu'elle remplit et certains aspects de sa structure, mais également un grand nombre de particularités lui donnent un statut à part. La plupart de ces particularités découlent de la différence de canal.

2.3.1 Paramètres impliqués

Une des conséquences de l'utilisation du canal visuo gestuel pour la communication en langue des signes est l'utilisation simultanée de nombreux paramètres.
Une première description des paramètres a été proposée par Stokoe [SCC78] pour établir un inventaire des signes de la Langue des Signes Américaine (ASL). Pour mettre en évidence les paramètres constitutifs des signes, Stokoe se base sur des paires minimales de signes ne différant que par un élément qu'il nomme *chérème*. Par analogie avec les *phonèmes* qui peuvent être assemblés pour former les signes, Stokoe distingue des *chérèmes* qui permettent de former des *kinèmes* (signes). Trois Chérèmes (emplacements, mouvements et configurations manuelles) réalisés simultanément (par opposition aux langues vocales plus

[1] Le concept d'iconicité sera expliqué §2.3.3

[2] Nous emploierons dans cet ouvrage la dénomination "langue orale" par opposition à la langue sous sa forme écrite. Le terme "langue vocale" sera réservé aux langues faisant un usage intensif de la voix.

séquentielles) permettent ainsi de réaliser les signes. Il dénombre en tout 12 emplacements, 24 mouvements et 19 configurations manuelles. Stokoe ajoute également dans sa description les notions de contact entre différents articulateurs et indique parfois l'orientation des mains sans leur prêter toutefois le statut de paramètres indépendants. De même, il repère l'expression du visage comme un élément linguistique sans y accorder la même importance qu'aux paramètres manuels. En montrant la décomposition de signes en chérèmes, Stokoe tente de démontrer la double articulation[3] de l'ASL, ce qui constitue une étape importante pour leur reconnaissance en tant que langue à part entière.

Par la suite, Battison [Bat74] mettra en évidence un quatrième paramètre d'orientation en utilisant ce même procédé de paires minimales. Cependant ce statut de paramètre autonome fera débat dans la communauté linguistique car certains linguistes estimeront que l'orientation est déductible des autres paramètres.

Plusieurs paires minimales permettent également de mettre en évidence le rôle important de l'expression du visage dans la reconnaissance des signes isolés. Ainsi, Baker [Bac76] cité par [Bou09], présentera ce paramètre comme paramètre mineur dans la productions de signes isolés.

Suite à l'ensemble de ces travaux, on distingue traditionnellement 5 paramètres dans les productions de langues des signes :

- La configuration manuelle
- Le mouvement manuel
- L'orientation manuelle
- L'emplacement des mains
- L'expression du visage

Lorsqu'on considère non plus le signe isolé, mais l'énoncé dans son intégralité, il est possible d'identifier de nombreux autres paramètres porteurs de sens. Les paramètres non-manuels ne sont alors pas réductibles à la seule expression du visage. Nous proposons donc de classer les paramètres de la manière suivante :

- Paramètres manuels
  - Orientation
  - Configuration
  - Emplacement
  - Mouvement
- Paramètres non manuels

[3] La double articulation mise en évidence pour la première fois par A. Martinet [Mar70] est la propriété de la langue d'être décomposée en unités discrètes à deux niveaux. La phrase peut être décomposée en monèmes (unités de signification minimale). Ces monèmes peuvent ensuite être décomposés en phonèmes qui ne sont cette fois-ci plus porteurs de sens.

- Buste, épaules, orientation de la tête, orientation de la tête, regard
- Lèvres, joues, ouverture des yeux, sourcils

Nous avons volontairement choisi de grouper les paramètres buste, épaules, orientation de la tête et orientation de la tête car ils remplissent des fonctions importantes dans la production de pointage, la localisation de nouveaux concepts dans l'espace et la prise de rôle [Meu08]. Le regard joue un rôle prépondérant dans la structuration d'un énoncé. Au contraire, les lèvres, les joues, l'ouverture des yeux, les sourcils ont à la fois une valeur modale (indiquant par exemple si l'énoncé est une interrogation, une exclamation, une affirmation) et une valeur aspectuelle (durée d'une action, quantification, détermination) [Cux00]. Les lèvres jouent également un rôle à part en permettant de labialiser un mot durant la production d'un signe.

Les différents paramètres que nous venons d'énumérer peuvent être impliqués simultanément ou de manière séquentielle dans la production d'énoncés. La mobilisation de ces paramètres permet également de transmettre parallèlement plusieurs informations.

2.3.2 Une organisation dans le temps et l'espace

Une autre conséquence de l'utilisation du canal visuo-gestuel pour la transmission du message est l'utilisation intensive de l'espace dans la structure d'énoncés en LSF. Ainsi, une production en LSF sera ordonnée à la fois dans le temps et l'espace. Les concepts utilisés lors de la production d'un énoncé peuvent être placés dans "l'espace de signation" qui peut être modélisé par un espace contenu dans une demi sphère située devant le signeur (voir figure 7.26).

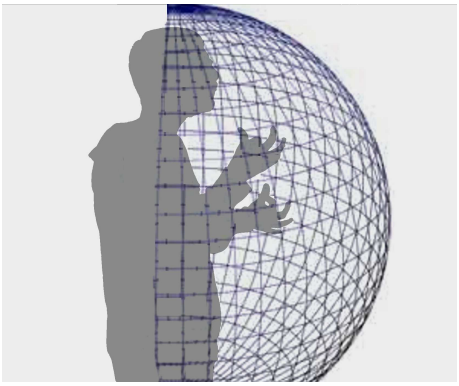


FIGURE 2.1 – Espace dans lequel sont couramment signés les signes

Une des conséquences de cette spatialisation des concepts est la flexion spatiale des signes. Ainsi, le verbe [DONNER] pourra être effectué avec des mouvements différents suivant les actants impliqués dans l'action. Il est possible d'identifier de grandes tendances dans la structure temporelle des énoncés signés. Pour n'en citer que quelques unes, on observe souvent une organisation temporelle d'un énoncé allant du global au détail, de l'inanimé à l'animé, dans l'ordre chronologique, positionnant le verbe en fin de phrase. Malgré tout, il ne s'agit pas de règles strictes d'ordonnancement temporel des concepts comme on pourrait en trouver dans les langues vocales.

2.3.3 L'iconicité

"L'iconicité est le lien de ressemblance plus ou moins étroit entre la chose du monde, le référent et le signe qui s'y rapporte" (Définition tirée [Cux00]). Les langues vocales n'y ont que rarement recours lors de la description du bruits des objets sous forme d'onomatopées (ex : Zzzzz pour imiter un moustique). On note toutefois parfois une utilisation des gestes coverbaux dans un but iconique comme "hier j'ai pêché un poisson gros comme ça" (où le geste indiquera la dimension du poisson).

En revanche, ce phénomène de "dire en montrant" est loin d'être marginal dans les Langues des Signes. Les premières études sur les LS ont bien souligné cette propriété sans pour autant lui accorder un statut linguistique, en préférant se focaliser sur la démonstration de la double articulation des langues des signes [SCC78]. Dans les années 90, Cuxac approfondira l'observation de l'iconicité et démontrera qu'il s'agit bel et bien d'un facteur structurant des Langues des Signes. Ainsi, il existe selon l'auteur deux manières de "dire" qu'il nomme visées iconiques : Il est possible de "dire sans montrer" ou de "dire en montrant". Les productions signées présentent de nombreux va-et-vient entre ces deux visées non-illustrative et illustrative.

La visée illustrative a été successivement étudiée par C. Cuxac et M. Sallandre. Cuxac distingue dans un premier temps trois structures de "grande iconicité" [Cux00].

- Les transferts de taille et de forme (TFF) permettent de donner à voir la forme d'un objet en la déployant dans l'espace. Certaines configurations manuelles nommées proformes permettent d'évoquer la forme de l'objet (voir 2.2). Cependant, les transferts de taille et de forme mobilisent également, en plus des paramètres manuels, un ensemble de paramètres non manuels indiquant la taille, le poids ou l'aspect de l'objet décrit. La figure 2.3 donne un exemple d'une mimique faciale soulignant l'immensité d'un immeuble.

- Les transferts situationnels (TS) permettent de placer des concepts dans l'espace de signation pour donner à voir leur positionnement relatif. Ces transferts peuvent également permettre de visualiser les déplacements d'une entité mobile dans l'espace de signation.

- Les transferts personnels (TP) permettent au signeur de prendre le rôle d'un des actants de son énoncé. Durant ces structures, le locuteur prend l'expression du visage et la manière de se comporter ou de signer de la personne dont il joue le rôle. Son orientation de buste et de regard est celle de la personne qu'il souhaite momentanément donner à voir [Meu08].

M. Sallandre précise ensuite ce modèle en distinguant d'autres structures de transfert qui sont pour la plupart des variantes ou des combinaisons des trois structures de grande iconicité que nous venons de mentionner. Plutôt que de détailler la vingtaine de variantes une à une, nous préférons mettre en évidence les différentes options dont dispose le signeur lors de la production d'un énoncé.

Il existe différents degrés de transferts personnels :

- Un transfert personnel de longue durée (par exemple dans le cas d'actions ou de discours rapportés).

FIGURE 2.2 – Utilisation d'un transfert de forme pour donner à voir la forme d'une balance (Vidéo tirée du corpus LS Colin [Cux02])

FIGURE 2.3 – Utilisation de la mimique faciale pour "qualifier" un signe (Vidéo tirée du corpus LS Colin [Cux02])

- Un semi-transfert ou transfert ponctuel (ceci est le cas lors d'une prise de rôle accompagnant la production d'un verbe spatialisé ou d'une action ponctuelle).
- Un pseudo transfert dans lequel la prise de rôle fait partie du signe (exemple : le signe [BARMAN]) mais ou le signeur ne joue pas à proprement parler le rôle de la personne.
- Naturellement, le signeur peut également ne pas prendre de rôle et n'être que narrateur extérieur.

[Sal03] observe que les transferts personnels peuvent être associés à des proformes (configurations manuelles particulières) évoquant une particularité physique du personnage (ex : les cornes d'une vache figure 2.4) ou soulignant une action du personnage (ex : regarder figure 2.5).

Lors de l'exécution d'un transfert situationnel, le signeur peut utiliser sa main dominée pour situer les autres signes par rapport à un référent stable figurant une entité de l'énoncé.

Les signes effectués par le signeur peuvent être de nature plus ou moins standard et appartiennent à plusieurs catégories :

- Le geste coverbal
- Le stéréotype : les gestes reflètent alors l'état d'esprit du personnage (ex : se gratter la tête lors d'une réflexion intense),

FIGURE 2.4 – Utilisation d'un proforme pour souligner une particularité du personnage lors d'un transfert personnel (Vidéo tirée du corpus LS Colin [Cux02])

FIGURE 2.5 – Utilisation d'un proforme pour souligner une action du personnage lors d'un transfert personnel (Vidéo tirée du corpus LS Colin [Cux02])

- Le signe de grande iconicité
- Le signe dérivé de l'iconicité standard utilisant des proformes.
- Le signe standard.

Il est souvent relativement difficile de déterminer si un signe relève de la grande iconicité, de l'iconicité standard (par le biais de proformes) ou est standard. En effet, il n'existe à l'heure actuel ni un recensement exhaustif des signes standards, ni une instance reconnue pour décider si un signe donné est standard ou non. Toutefois, le fait que le signeur regarde un ses mains lors de la production d'un signe lui confère souvent une valeur de signe de grande iconicité. Une des raisons de la difficulté à établir une liste de signes standards est le continuel enrichissement de la LSF par de nouveaux signes.

L'intérêt d'un travail tel que celui de M. Sallandre réside avant tout dans le fait de montrer qu'il peut y avoir plusieurs simultanéités dans une production en LSF :

- Une simultanéité de plusieurs informations pouvant être véhiculées par des signes standards ou non standards (issus de structures de grande iconicité ou d'utilisation de proformes),
- Une simultanéité de plusieurs visées pouvant être simultanément illustratives et non-illustratives.

Le processus de création de nouveaux signes (néologismes) dont nous avons discuté plus haut est aussi guidé par l'iconicité. Il suffit pour s'en convaincre de regarder l'ouvrage [Del07] sur l'étymologie des signes standards. Pour Cuxac, le processus de création de signe s'appuie d'abord sur une composition de morphèmes (proformes, mouvements ou emplacements signifiants). Ensuite, le signe évolue en raison de contraintes physiologiques et perceptives. Il en résulte une simplification du signe qui s'effectue tout en gardant la racine iconique du signe. [DGTC08] ont également observé ce processus de création lexicale dans un contexte de cours de langue des signes. L'étude confirme que les néologismes et les propositions de signes sont issus d'une lexicalisation de transferts de taille et de forme, de transferts situationnels ou d'une utilisation de proformes. Ce processus de création de signes basé sur des morphèmes (unités minimales de sens) permet de créer une représentation des concepts logique et souvent aisément compréhensible par un interlocuteur connaissant les structures de grande iconicité de la LSF.

2.3.4 Conséquences pour un traitement automatique de la LSF

Toutes ces particularités ont naturellement d'énormes répercutions dans le domaine de traitement automatique des langues des signes. Outre les problèmes d'acquisition des valeurs des paramètres à partir de la vidéo, de nombreuses problématiques sont à surmonter pour traiter des énoncés représentatifs de la LSF :

- Il s'agit tout d'abord de pouvoir intégrer dans un même système des données extrêmement hétérogènes (ex : la configuration manuelle et la rotation du buste).
- Un même paramètre doit pouvoir être interprété à différents niveaux. Si on prend l'exemple du mouvement, celui ci peut être interprété :

  Au niveau sémantique Dans le cadre de néologismes, le mouvement peut directement constituer un morphème participant au sens du signe créé[4]. De plus, la modification du mouvement d'un signe standard (plus grande amplitude, répétition, vitesse d'effectuation) est également porteuse de sens.

  Au niveau lexical Le mouvement fait partie de la description du signe standard. Nous considérons que le fait qu'une main reste statique pendant l'effectuation d'un signe peut également être considéré comme un mouvement qui peut avoir du sens.

  Au niveau syntaxique Le mouvement permet d'accorder certains verbes dans l'espace (c'est à dire d'indiquer quels actants ou entités ils impliquent. On pourrait faire un parallèle avec l'accord d'un verbe avec son sujet et ses compléments pour les langues vocales).

[4] On donnera l'exemple du signe [DONNER] impliquant deux actants A et B ou le sens du mouvement permet de montrer si A donne à B ou B donne à A

Au niveau prosodique Certains attributs du mouvement (tonicité, amplitude, vitesse d'effectuation) sont également modifiés de manière suprasegmentale en fonction du locuteur, de la personne transférée ou de la modalité de l'énoncé.

- Le système doit pouvoir intégrer à la fois les signes non standards et les néologismes très fréquents en Langue des Signes.
- Les signes standards peuvent être réalisés avec une énorme variabilité.
- Les structures de grande iconicité donnent à voir plus directement le sens de l'énoncé. Leur interprétation automatique est probablement du même ordre de difficulté que l'interprétation automatique d'images.
- Il est possible de transmettre de manière simultanée plusieurs informations sous plusieurs visées différentes (par exemple, dans le cas de double transfert).

Les défis à surmonter dans le domaine du traitement automatique sont donc immenses et le problème offre un complexité tout à fait différente de celle des langues vocales. Le fait d'avoir objectivé l'ensemble de ces contraintes nous guidera lors de la recherche de solutions pour le traitement automatique des LS et nous y reviendrons lors de l'évaluation des solutions que nous proposons.

CHAPITRE 3

LA NOTATION DU SIGNE

Il est possible d'aborder le Traitement Automatique des Langues des Signes (TALS) à plusieurs niveaux. Nous avons décidé de nous intéresser dans ce travail aux mouvements manuels dans la production de signe car nous disposons d'assez d'outils de traitement d'image pour exploiter ce paramètre. Nous choisissons donc d'écarter volontairement de notre analyse les problèmes relatifs à la structure des énoncés signés. Nous y ferons cependant référence ponctuellement car la structure spatiale de la langue des signes a de nombreuses répercussions sur la structure des signes.

Avant de détailler les différentes manières de modéliser les unités gestuelles impliquées dans la production d'énoncés signés, il nous paraît nécessaire de nous attarder sur la notion de signe. Ce concept de "signe" revêt en effet des acceptions bien différentes suivant qu'on se place dans la perspective d'un philosophe, d'un linguiste et suivant qu'on étudie une langue vocale ou signée.

Par conséquent, nous proposons dans un premier temps une définition philosophique du signe que nous appliquons successivement aux langues vocales et signées. Ceci nous permet de mettre en évidence des différences fondamentales entre les systèmes de codages impliqués dans ces deux modes de communication. Nous aboutirons également à des définitions plus précises des concepts de "signes standards" et "d'unités gestuelles iconiques" que nous utiliserons dans la suite de ce manuscrit.

3.1 Qu'est ce qu'un signe

Avant d'entrer dans la définition d'un signe dans le cadre des langues, attardons nous sur la définition philosophique qu'on peut donner de ce concept. Une réponse nous est proposée par U. Eco dans son ouvrage sur "le Signe" [Eco88]. L'auteur dégage trois composantes du signe :

Le signifié est le concept représenté par le signe. Le signifié est une notion abstraite, idéalisée.

Le référent est un objet du monde réel appartenant à la catégorie du signifié.

Le signifiant est la représentation artificielle et codée du signifié (sous forme de son, de mot, de dessin, de symbole ...).

Prenons l'exemple concret du mot "arbre" et regardons comment il peut être décliné sous plusieurs facettes :

Signifié Il s'agit là du concept d'arbre : "Végétal dont la tige, appelée fût ou tronc, ne se garnit de branches et de feuilles qu'à une certaine hauteur" [1].

[1] définition tirée du site fr.wiktionary.org/wiki/arbre

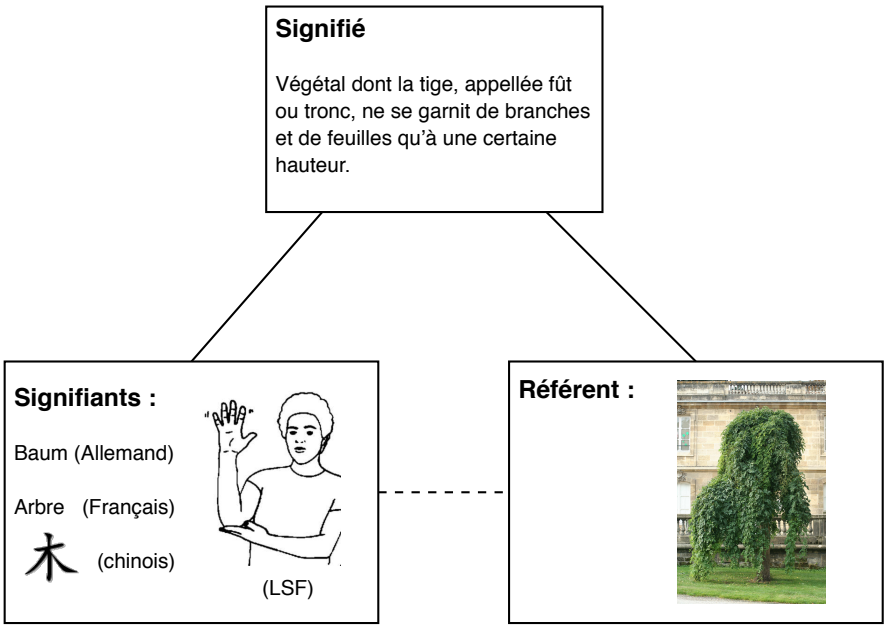


FIGURE 3.1 – Exemple d'association signifiant / signifié / référent

Le référent est un arbre précis, par exemple celui présent dans la figure 3.1 [2].

Le signifiant est une représentation symbolique du concept. La figure 3.1 en propose plusieurs sous forme de pictogrammes, de Sinogramme, de mots écrits sous forme alphabétiques, de geste [3].

Il est important de noter que, suivant le codage retenu, un signifiant différent est choisi.

3.1.1 Réponse de la linguistique générale

Cette notion de signe peut naturellement être appliquée à la linguistique des langues vocales. Sans développer les différentes approches pour définir le signe linguistique, nous nous baserons dans un premier temps sur l'analyse de F. de Saussure [Sau16]. Pour l'auteur, le signe linguistique possède quatre caractéristiques significatives :

L'arbitraire Le lien entre signifié et signifiant est arbitraire. C'est à dire que la relation entre l'image accoustique et le concept désigné est fixée arbitrairement pour chaque langue. De Saussure ira même jusqu'à dire que cette relation n'est pas motivée[4].

Le caractère linéaire du signifiant Le signifiant des langues vocales étant de nature linéaire (texte écrit, image acoustique), les constituants du signifiant sont placés obligatoirement les uns après les autres.

[2] Image d'arbre tirée du site www.sharkdesign.fr

[3] Illustration tirée de [MVGD97a]

[4] "Le mot arbitraire appelle aussi une remarque. Il ne doit pas donner l'idée que le signifiant dépend du libre choix du sujet parlant (on verra plus bas qu'il n'est pas au pouvoir de l'individu de rien changer à un signe une fois établi dans un groupe linguistique) ; nous voulons dire qu'il est immotivé, c'est-à-dire arbitraire par rapport au signifié, avec lequel il n'a aucune attache naturelle dans la réalité." [Sau16]

L'immutabilité synchronique Un locuteur ne peut pas modifier arbitrairement un signe car le signifiant associé au signifié est commun à une communauté linguistique.

La mutabilité diachronique Les signes linguistiques peuvent admettre des modifications au cours du temps.

Si les trois derniers points sont admis par la communauté des linguistes, le premier fait toujours débat. La remise en cause porte en particulier sur le caractère non motivé du signe linguistique. Comme l'a montré le linguiste [Jes22], il est possible de trouver une sorte de symbolisme phonétique dans certains mots de la langue anglaise. Il s'agit donc pour l'auteur de mettre en évidence une certaine iconicité des langues vocales dont un relicat se retrouverait dans les mots actuels.

Il est intéressant de mentionner également la manière dont Hjemslev appréhende le signe [Hje43]. Le linguiste distingue deux aspects du signe : l'expression (qu'on pourrait rapprocher de la notion de signifiant) et le contenu (équivalent du signifié). Le contenu et l'expression sont alors chacun divisé en substance et en forme. Les substances sont des espaces continus tandis que les formes sont des discrétisations de ces espaces en classes distinctes.

Au niveau de l'expression Le message est véhiculé par un continuum de sons (qui constituent la substance de l'expression) qui peuvent être décrits grâce à la phonétique. Chaque langue possède ensuite sa propre discrétisation des sons en phonèmes et ses règles de combinaisons qui peuvent être décrites par la phonologie. Ces phonèmes combinés entre eux constituent la forme de l'expression.

Au niveau du contenu Il existe un continuum de concepts qui constituent la substance du contenu. Chaque langue choisit de discrétiser cet espace en plusieurs classes qui constituent la forme du contenu. Il est intéressant de noter que là aussi, la forme du contenu varie d'une langue à l'autre. L'illustration 3.2 montre une partition différente des couleurs suivant les cultures (preuve qu'il peut exister plusieurs formes pour une même substance).

Le signe consiste à une association entre deux ensembles de ces espaces (fig. 3.3).

FIGURE 3.2 – Exemple de découpage différents d'un même spectre pour les langues anglaise et galloise

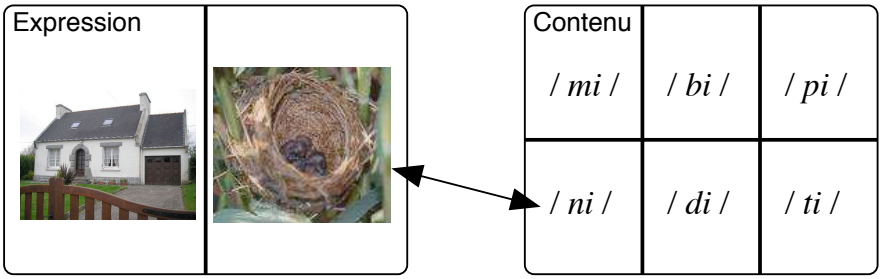


FIGURE 3.3 – Exemple d'association entre l'espace des contenus et l'espace des expressions dans le cas des langues vocales

	ovin	porcin	bovin	équin	humain
mâle	mouton	porc	taureau	étalon	homme
femelle	brebis	truie	vache	jument	femme

FIGURE 3.4 – Exemple de décomposition du sens de signes tiré de [Hje43]

Ayant effectué ce parallélisme de structure forme/substance entre les niveaux contenu et expression, Hjemslev y applique les mêmes outils et décompose le sens des signes en composants de base qu'il nomme sémème. Le tableau 3.4 montre comment il décompose le sens des désignations des animaux. Bien qu'on puisse remettre en cause cette décomposition en élément de sens minimaux pour tous les concepts[5], il est important de souligner cette démarche de discrétisation des unités de sens.

Reste qu'une fois le terme de signe ainsi défini, il est difficile de proposer une segmentation d'un énoncé d'une langue vocale en signes. Une solution de facilité pourrait être de considérer le "mot"[6] comme équivalent du signe linguistique. Il suffirait alors de considérer les signes comme des suites de caractères précédées et suivies par des signes de ponctuation ou des espaces. Cette solution est loin d'être satisfaisante comme le montrent les observations suivantes :

- Lors de l'utilisation d'une langue vocale, l'espace présent dans l'écriture n'est pas forcément audible sous forme d'une pause.
- Le découpage des mots est variable en fonction du genre et du nombre (ex : "ceci", "ceux-ci").
- Certaines langues comme le chinois ne séparent pas les mots par un espace.
- Qu'en est-il de tous les mots composés ? (par exemple, une porte-voix doit il être considéré comme un seul mot ou comme deux mots ?)

[5]On pourra se reporter en particulier à l'ouvrage [Cho57] pour illustrer la limite de ce procédé de décomposition du sens des mots.

[6]Etant donné la nature de notre travail de recherche, nous ne rentrerons pas en détail dans la définition des notions de morphème ou de lemme qui seraient plus précises linguistiquement.

Une solution alternative serait de considérer le signe comme la plus petite unité de sens qu'il est possible d'isoler dans un énoncé. Un des problème de cette approche est qu'il est fréquent qu'une unité de sens soit fractionnée (morphème à signifiant discontinu). Par exemple, dans la phrase "il ne veut pas", il est possible d'identifier "ne ... pas" comme unité de sens minimale. Malgré la définition théorique du signe linguistique, il est donc difficile de décomposer un énoncé (qu'il soit écrit ou oral) en signes.

3.1.2 Réponse spécifique aux langues des signes

Avant de tenter de définir un signe dans le cas des Langues des Signes, revenons sur les différentes catégories dans lesquelles il est possible de classer les unités gestuelles. Nous nous basons sur l'analyse de Kendon [Ken88] cité par [McN05] qui situe les gestes communicatifs sur un continuum :

- Le premier stade est désigné par le terme **gesticulation**. Il s'agit d'un ensemble de gestes coverbaux (accompagnant la parole) comprenant des pointages déictiques, des gestes iconiques et des gestes marquant le rythme du discours.
- Ensuite, viennent ce que l'auteur appelle **gestes de para-langage**. Il s'agit d'un ensemble réduit de gestes conventionnels comme ceux utilisés dans le cadre d'arbitrages sportifs. Il n'est pas en revanche possible de combiner ces gestes pour produire un énoncé plus complexe.
- Vient ensuite le **pantomime** (ou mime).
- Les **emblèmes** ou gestes symboliques peuvent être utilisés en complément ou en parallèle des mots articulés. Ces codes gestuels sont partagés au sein d'une société et peuvent donc différer d'un pays à l'autre. Une bonne illustration de différence d'emblème d'une culture à une autre est la différence entre le système pour compter sur ses doigts en France et en Chine (où les configurations manuelles reprennent de manière iconique la forme des caractères chinois correspondants).
- Les **Langues des signes** sont les seules à pouvoir être considérées comme langues à part entière.

Les Langues des Signes n'excluent cependant pas l'utilisation des emblèmes, du mime, des gestes de para-language et des gestes coverbaux comme les pointages et les gestes iconiques qui constituent une pierre d'angle de la langue des signes (cf. §2.3.3). Ce qu'il est important de retenir, c'est que la notion de signe dans le cadre de la Langue des Signes recouvre par conséquent des réalités extrêmement différentes suivant le signe considéré.

Si on revient sur la caractérisation que propose F. De Saussure (cf. §3.1.1), deux aspects diffèrent des langues vocales. Le premier concerne naturellement le caractère immotivé du signe. Comme l'a montré Cuxac [Cux00], la motivation des unités gestuelles transparaît au travers de leur iconicité et des proformes qui les composent. Cette iconicité est la condition pour pouvoir réaliser un va et vient fréquent entre les visées illustratives et non-illustratives. Ceci ne remet en aucun cas en cause le caractère conventionnel des

signes standards. En d'autres termes, les traits constitutifs des signes remplissent à la fois une fonction différentielle et une fonction référentielle. La deuxième grande différence concerne la linéarité des unités gestuelles. Contrairement aux mots des langues vocales qui peuvent être considérés comme des signaux à une dimension sonore, la production d'une unité gestuelle de Langue des Signes met en oeuvre l'utilisation parfois simultanée d'une dizaine de paramètres qui varient au cours du temps.

Prenons cette fois la définition proposée par Hjelmslev et tentons de l'appliquer aux langues des signes (nous nous limiterons dans notre analyse à l'étude des signes standards). Plusieurs éléments diffèrent avec les langues vocales.
D'une part, les différents paramètres impliqués dans la production d'un signe ne sont certainement pas tous discrets. La continuité semble vérifiée pour des paramètres comme l'amplitude et la vitesse des mouvements. La thèse de L. Boutora [Bou09] ne conclut pas non plus clairement sur la perception catégorielle des configurations manuelles.
D'autre part, la discrétisation des concepts prend ici un sens un peu différent. La forme du contenu du signe est susceptible de varier significativement en fonction des paramètres de réalisation de l'unité gestuelle.

Nous pouvons maintenant reprendre l'idée d'Hjelmslev qu'il existe un homomorphisme entre espace des expressions et espace des contenus et formuler une hypothèse[7]. L'iconicité des Langues des Signes permet d'aller plus loin dans l'association entre expression et contenu. Pour chaque signe, il sera également possible de proposer une corrélation entre des variations des paramètres constitutifs des unités gestuelles et des variations d'éléments constitutifs des signes.

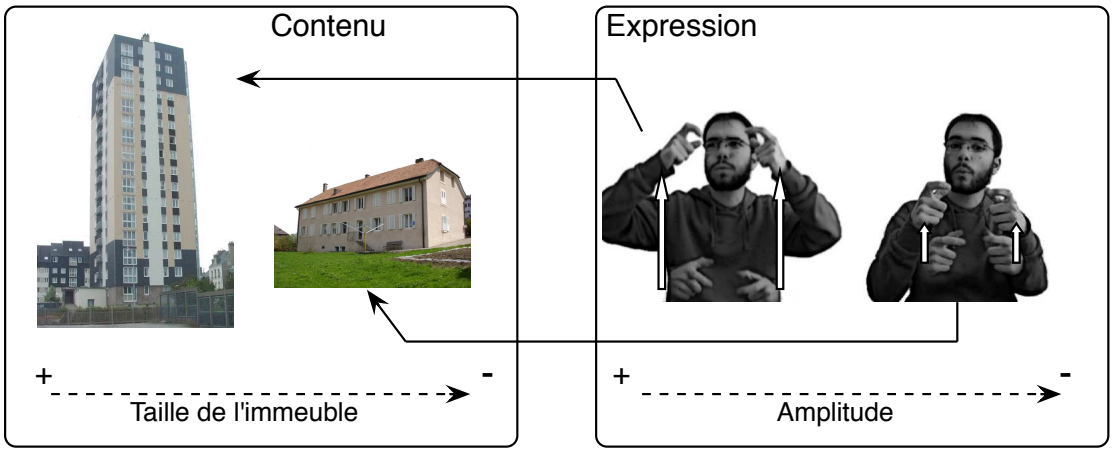


FIGURE 3.5 – Correspondance entre la variation de la forme et la variation de l'expression dans le cas d'unités gestuelles

Donnons une illustration de cette hypothèse avec le signe immeuble(figure 3.5) :

[7]Il est toutefois aussi possible de définir l'iconicité comme une analogie [Eco88]

- L'espacement entre les deux mains du signeur peut être associée à la longueur de la barre d'immeuble dont on parle.
- L'espacement entre le pouce et l'index du signeur peut indiquer la largeur de la barre d'immeuble.
- L'emplacement du signe indique, dans le cas d'un transfert situationnel, la place réelle de l'immeuble.
- La direction du mouvement indiquera si la tour est verticale ou penchée.
- L'amplitude du mouvement, sa vitesse d'effectuation ainsi que l'expression du visage peut indiquer la hauteur de l'immeuble.
- L'expression du visage peut indiquer l'aspect de l'immeuble.

Comme nous venons de le voir, la nature des unités gestuelles est extrêmement différente de celle des mots des langues vocales. Il faut traiter en langue des signes, des unités gestuelles dont la grande variabilité des paramètres peut être signifiante.

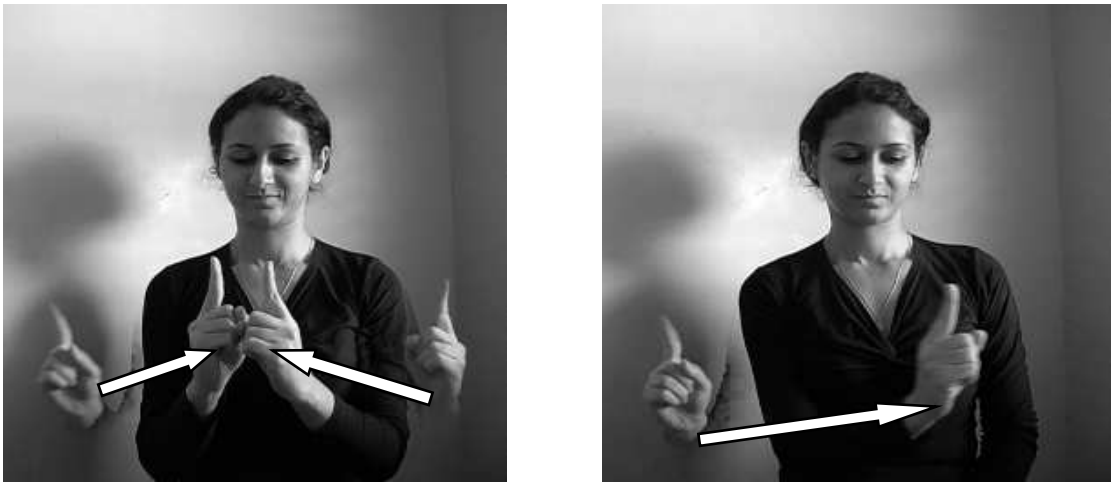


FIGURE 3.6 – Signes [RENCONTRER] et [UNE PERSONNE PASSE]

Est-il pour autant possible de définir ce qu'est un signe avec précision ? La question est aussi délicate que pour les langues vocales. Il n'est pas possible d'utiliser l'écriture des LS pour observer d'éventuels espaces car il n'existe pas encore d'écriture standardisée. Le problème de considérer un signe comme une unité minimale de sens est aussi problématique. En effet, il est souvent possible de décomposer un signe standard en plusieurs sous parties, elles-mêmes porteuses de sens. Pour s'en convaincre, on peut considérer l'exemple du signe [RENCONTRER] qui peut être vu comme une concaténation de deux exemplaires du signe [UNE PERSONNE PASSE] (cf fig. 3.6).
On retrouve également l'épineux problème des signes composés. Ces signes sont issus de la juxtaposition

temporelle de deux signes standard ; les nouveaux signes prennent alors des sens différents. A partir de quand peut on considérer que deux signes forment un signe composé ? Encore une fois, il n'existe pas encore de réponse définitive à cette question.

Nous n'avons jusqu'ici abordé que la description les signes standards. Il faut savoir qu'il existe en réalité de nombreuses "unités gestuelles" non-standard qui représentent pourtant une part importante des productions en Langues des Signes [Sal03]. Ces unités gestuelles impliquent souvent beaucoup les paramètres non manuels.

3.1.3 Pertinence de la notion de signe

Comme nous l'avons montré dans les parties précédentes, la notion de signe, même si elle est définie philosophiquement est difficilement utilisable pour mener à bien une décomposition d'un énoncé en signes qui ne soit pas sujette à discussion. Pour autant, l'étude des types de signes, de la segmentation d'un énoncé en signe et de la représentation graphique du signe est-elle pertinente ? Nous répondons par l'affirmative en nous basant sur les faits suivants :

- Si on demande à plusieurs sujets maîtrisant la LSF de segmenter un énoncé sous forme d'une succession de signes et de transitions (intervalle temporel entre deux signes consécutifs), on note que les segmentations sont compatibles (cf. §8.3).
- Même si la notion de signe n'est pas clairement définie, on assiste depuis quelques années dans la communauté des linguistes, l'émergence de plusieurs travaux qui s'attachent à décrire la phonologie et à la morphologie des signes [Bou09] [Bou07].
- Il existe des dictionnaires dans lesquels les signes standards sont décrits [MVGD97a] [Com06]. Il est même possible d'effectuer des recherches de ces signes en utilisant leurs paramètres (configuration, emplacement, mouvement) [Fou07].

Cependant, nous notons que ce découpage en signe est clair uniquement pour les signes standards et qu'il est beaucoup plus variable pour l'ensemble des unités gestuelles iconiques (il n'est d'ailleurs pas rare qu'une personne propose deux segmentations différentes d'un même énoncé à quelques jours d'intervalle pour des unités gestuelles participant à un transfert).

3.1.4 Bilan

Comme nous l'avons vu dans l'analyse précédente, la notion de signe a pu être définie philosophiquement mais il est difficile d'appliquer les définitions théoriques pour expliquer ce qu'est précisément un signe d'une langue vocale ou signée. Malgré tout, un certain nombre d'observations et d'expérimentations nous permettent de valider empiriquement la pertinence de la notion de signe. Dans la suite de ce travail, nous emploierons le mot "signe" pour désigner une unité gestuelle reconnue comme signe standard par un

certain nombre de locuteurs de la LSF. Nous utiliserons également la désignation “signe composé” pour les signes formés d’une concaténation de signes standards dont le sens est différent de la juxtaposition des sens des deux signes d’origine. Les unités gestuelles utilisées dans le cadre de transferts seront appelées “unités gestuelles de grande iconicité”.

3.2 La description des signes

Après avoir précisé la notion de signe en LS, il est naturel de se demander la manière d’en donner une description pertinente d’un point de vue graphique ou informatique, de façon à pouvoir mener ultérieurement un traitement automatique. Un premier point mérite d’être souligné : les langues des signes sont des langues orales qui n’admettent pas encore d’écritures uniformisées. Nous allons donc décrire plusieurs formalismes de notation qui ont été mis au point pour garder une trace écrite ou informatique des signes, en mettant en évidence les forces et les faiblesses des différentes méthodes et en faisant le lien avec les perspectives que chacune d’elles offre pour le traitement automatique des langues des signes. Pour des raison de concision, nous n’aborderons pas les modèles de signes tels que ceux décrits par Bebian [BA25], Laban [Lab74] et Jouison [Jou95] dont nous ne faisons pas usage dans notre travail de recherche.

Du point de vue informatique, nous ne mentionnerons que les modèles utilisés dans un but de génération d’énoncés. Les modèles d’analyse seront abordés dans le chapitre 5 sur la caractérisation des signes. Le lecteur qui souhaiterait en savoir plus sur les spécificités des différentes méthodes de notations graphiques du signe pourra se reporter à [Bou09].

3.2.1 Approches explicites

La première approche pour garder une trace d’une production en langue des signes est de la filmer. Cette méthode d’écriture-vidéographique que nous nommerons LS-Video permet de préserver les spécificités des langues des signes (spatialisation, parallélisme des information, continuité des phonèmes, richesse des paramètres non-manuels). Elle partage en outre de nombreux points communs avec l’écrit :

- La vidéo permet de transmettre le message qui peut être reçu en différé par un ou plusieurs destinataires. Elle peut également être reproduite.
- Le discours peut aussi faire l’objet d’une autocorrection par le locuteur.
- De la même manière qu’il est possible de réorganiser un texte, d’y ajouter des notes de bas de page, d’en tirer des extraits, il est possible d’effectuer des montages vidéos effectuant des fonctions similaires [LADD$^+$re].

Il est toutefois nécessaire de souligner quelques différences fondamentales entre l’écriture et la vidéo. La vidéo ne préserve pas l’anonymat du locuteur. On peut également objecter le fait que la vidéo ne correspond pas aux critères de caractérisation des écritures conventionnelles et ne constitue pas un “Système de signes

graphique servant à noter la parole ou la pensée"[8]. En effet, le système d'enregistrement est dépourvu de tout symbolisme et ne permet pas de reproduction graphique (en 2D) facile à comprendre.

Une des solutions pour conserver l'anonymat du locuteur peut être d'enregistrer sa production par le biais de dispositifs de capture de mouvement (voir chapitre 4) ou bien de reproduire les mouvements du locuteur par des techniques comme la rotoscopie [BBFV07]. Ces deux techniques permettent de conserver l'expressivité du signeur et d'effectuer le rendu des mouvements sur un signeur virtuel. L'anonymat du locuteur n'est en fait que partiellement préservé. Comme l'ont montré les expérimentations menées dans l'entreprise Websourd, il est possible de reconnaître le style du signeur grâce à la prosodie dont sont empreints les mouvements reproduits. Il serait donc possible de reconnaître malgré tout la "voix" du signeur.

Même si les deux techniques de notation des signes que nous avons mentionnées permettent une reproduction très fidèle des LS, elles sont difficilement exploitables telles qu'elles dans le cadre du traitement automatique des LS. En effet, les enregistrements produits possèdent de nombreuses redondances et il n'est fait aucune sélection entre les informations pertinentes et celles qui sont non pertinentes du point de vue de la communication.

3.2.2 Notations symboliques

Pour les raisons que nous venons d'évoquer plus haut, nous nous intéresserons davantage dans cet état de l'art aux notations plus symboliques des LS qui sont exploitables d'un point de vue informatique. Nous écartons d'emblée la notation des signes sous forme de glose (traduction signe à mot d'une langue vocale) pour plusieurs raisons :

- Le fait de noter un signe comme chaîne de caractères est extrêmement réducteur et supprime toutes les nuances que nous avons pointées en 3.1.2. On supprime également toute la spatialité de la langue des signes.

- Il n'existe pas de bijection entre l'espace des signes d'une langue des signes (même pour les signes standards) et l'ensemble des mots d'une langue vocale.

- Les unités gestuelles de grande iconicité ne peuvent pas être retranscrites aisément par des mots.

D'un point de vue de traiteur d'image, la glose ne présente qu'un intérêt mineur dans la mesure où elle ne propose aucune formalisation du signe utilisable en traitement automatique.

Nous nous focalisons donc dans la suite de cet état de l'art sur les différentes méthodes permettant de représenter de manière symbolique et synthétique la manière dont sont réalisées les unités gestuelles impliquées dans la réalisation d'énoncés en LS. Nous exposons successivement les différentes approches puis nous proposons une synthèse en situant notre proposition par rapport aux modèles existants.

[8] Extrait de la définition du mot "écriture" tirée du dictionnaire Larousse Pratique (2005)

3.2.3 Approches paramétriques

Comme nous l'avons mentionné en 2.3.1. Stokoe a été un des premiers linguistes à proposer une décomposition des signes en différents paramètres. Le formalisme de notation qu'il propose spécifie pour chaque signe :

Des emplacements (Tab) Différents symboles graphiques permettent de spécifier en tout 12 emplacements de l'espace dans lesquels sont effectués les signes. Si la discrétisation des emplacements à proximité du visage est relativement fine, on note que tous les signes effectués devant le buste du signeur sont notés comme étant effectués dans l'"espace neutre" sans ajouter plus de précisions. Ceci pose notamment des problèmes pour indiquer la spatialisation des signes.

Des configurations manuelles (Dez) L'auteur distingue 19 configurations manuelles. L'ajout de symboles diacritiques permet éventuellement d'ajouter des variantes à partir de ces configurations de base. Il est possible de préciser éventuellement l'orientation de la paume de la main à partir d'autres symboles.

Des orientations (Sig) Le formalisme permet de représenter 24 mouvements de base impliquant une ou deux mains. Ces symboles de mouvements peuvent être combinés de manière à spécifier des mouvements plus complexes ou des mouvements simultanés des mains droite et gauche.

Des symboles additionnels permettent, si besoin, de spécifier la position relative des deux mains. La description de signe se fait en spécifiant les paramètres dans l'ordre [Tab][Dez][Sig] (voir fig. 3.7) Ce formalisme a été utilisé par Stokoe pour représenter les signes dans un dictionnaire anglais / ASL [SCC78]. Par contre, il ne spécifie que les paramètres manuels et est relativement imprécis pour indiquer la spatialisation du signe.

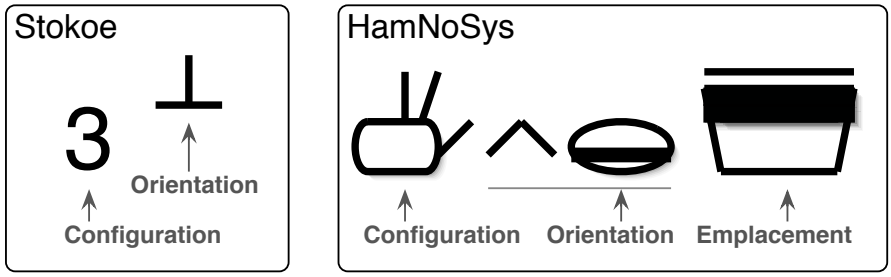


FIGURE 3.7 – Signe [TROIS] représenté à l'aide de la notation de Stokoe et du formalisme HamNoSys

Le système de notation HamNoSys (Hamburg Notation System) [Pri89] reprend les paramètres identifiés par Stokoe et enrichit le formalisme par quelques symboles indiquant des paramètres non-manuels. Les symboles retenus pour la description de l'emplacement, de la configuration, de l'orientation et du mouvement sont plus iconiques et permettent donc un déchiffrement plus rapide des signes par un lecteur. L'ajout de symboles spécifiant les contacts, les positions relatives des mains, les relations de symétrie et les répétitions de primitives de mouvement permettent une notation relativement synthétique des signes. En

raison du nombre important de symboles utilisés (on en recense environ 500), il est relativement difficile de les mémoriser tous. Bien que les valeurs des différents paramètres soient discrètes, il est possible d'indiquer des valeurs intermédiaires en effectuant des moyennes entre plusieurs valeurs discrètes.
Un logiciel nommé Ilex[9] permet de traiter informatiquement les signes décrits avec les symboles du formalisme et de restituer les signes à l'aide de signeurs virtuels. Le formalisme a été utilisé dans le cadre de traitement automatique des langues des signes dans les projets européens Visicast[10], eSign[11] et Dictasign[12].

3.2.4 Approches temporelles

Par opposition aux approches précédentes qui mettent l'accent sur la simultanéité de production des différents phonèmes, d'autres formalismes de notation des signes décomposent le signe en une séquence d'unités temporelles. Lidell et Johson [LJ90] distinguent ainsi deux types de phases :

Les Postures P (autrefois H) correspondent à des phases durant lesquelles les différents paramètres impliqués dans la production du signe ont une variation faible ou nulle. Le formalisme permet de retranscrire avec précision les emplacements (incluant les contacts), les configurations et orientations des mains durant les phases de tenues. Il est aussi possible d'indiquer la durée de ces phases.

Les Transitions T (autrefois M) correspondent à des phases durant lesquelles chaque paramètre est susceptible de varier. Il contient la forme des déplacements des mains (rectiligne, arrondi, avec un mouvement secondaire ...), le contact prolongé des mains, la dynamique des mouvements (balistiques ou tendus).

On aboutit donc ici à une description du signe en partition ou chaque ligne correspond à untrait phonétique [13] (emplacement, orientation, configuration, contact ...) et où les colonnes correspondent aux différentes phases de postures et de transitions (fig. 3.8) [Joh08] .

D'autres modèles que nous ne détaillerons pas pour des raisons de concision reprennent également cette notion de décomposition des signes en différentes phases. Ainsi Brentari [Bre98] dissocie les paramètres invariants (Inherent features) des paramètres variables (prosodic features) durant la production du signe et décompose les variations complexes des paramètres en plusieurs unités temporelles. Les modèles proposés par Perlmutter [Per92] et Sandler [San89] (pour n'en citer que quelques uns) peuvent aussi être classés dans cette catégorie d'approche temporelle de description des signes.

[9] http ://www.sign-lang.uni-hamburg.de/ilex/
[10] http ://www.visicast.co.uk/
[11] http ://www.sign-lang.uni-hamburg.de/esign/
[12] http ://www.dictasign.eu/
[13] La granularité de la description peut être adaptée suivant l'objectif recherché.

Phase temporelle		*posture*	*transition*	...	...
Main droite	Configuration				
	Orientation				
	Placement				
Main gauche	Configuration				
	Orientation				
	Placement				
Non manuel					

FIGURE 3.8 – Structure simplifiée de la description d'un signe à l'aide du formalisme de Johnson

Ces approches de description d'un signe sous forme multi-segmentale ont été adaptées dans le cadre de génération d'énoncés par des signeurs virtuels. Ils proposent en effet une description des signes compatible avec les méthodes d'animation des signeurs virtuels par images clés et interpolations. De plus, ces méthodes de description permettent de décrire n'importe quelle unité gestuelle manuelle.
Ainsi, O. Losson [Los00] propose un système de description informatique des signes proche de celui de Liddell et Johnson. Le formalisme retenu permet de spécifier des flexions sur les signes standards pour permettre de modifier plusieurs caractéristiques du signe comme son amplitude, sa dynamique ou son emplacement. Un autre formalisme de description des signes proposé dans [Fil08] se place également dans une logique multi-segmentale et insiste davantage sur les relations géométriques entre les différents paramètres manuels lors de la réalisation du signe. Les "degrés de liberté" du signe susceptibles de varier et d'être exploités iconiquement sont ainsi visibles dès la spécification du signe. L'illustration 3.9 permet par exemple de visualiser la spécification géométrique du signe [BALLON][14].

3.2.5 Description unisegmentale des signes

En opposition à ces modèles décrivant les signes comme une succession de segments temporels, un certain nombre de modèles décrivent un signe comme un unique segment. Ainsi, pour Van der Hulst [HVd93] cité par [Bou09] , "le mouvement n'est pas considéré comme unité phonologique, car il est déductible des spécifications de l'aperture de la main[15] pour le mouvement local, et des emplacements majeurs et mineurs pour le mouvement et le déplacement. Dans son sillage, Channon propose une modélisation du signe sous forme d'un segment unique [Cha02]. Outre le fait que le formalisme que l'auteur propose permet de représenter la majorité des signes de l'ASL de manière extrêmement synthétique, il est intéressant de relever son excellent argumentaire plaidant pour une description des signes sous forme d'un segment unique

[14]Illustration fournie par le LIMSI avec l'aimable autorisation de M. Filhol et A. Braffort
[15]orientation et configuration manuelle

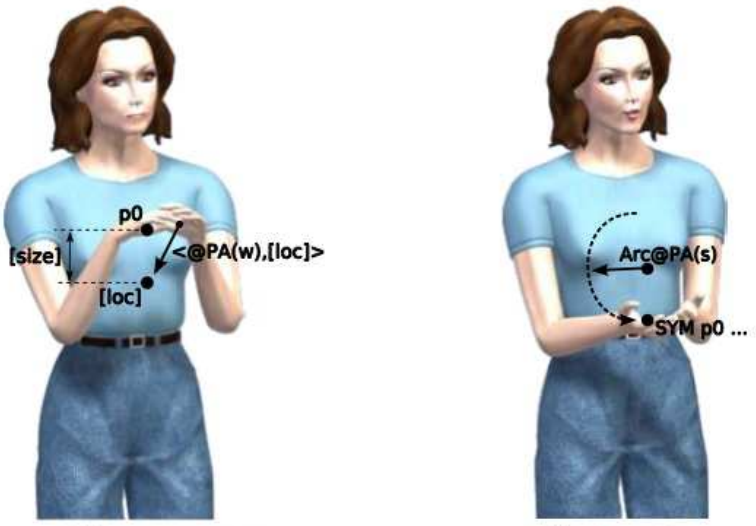


FIGURE 3.9 – Signe [BALLON] représenté à l'aide du formalisme Zebedee

basé sur la fréquence des répétitions et des symétries, ainsi que la redondance des approches de description temporelle des signes. Nous reprenons d'ailleurs la majorité de ces arguments à notre compte.

3.2.6 Autres approches

D'autres approches de représentation du signe se distinguent par l'utilisation de l'espace de leur support par analogie avec l'espace réel. Il s'agit pour toutes les méthodes de notation des signes que nous allons présenter, de formalismes graphiques à deux dimensions, par opposition à toutes les méthodes permettant de représenter un signe de manière linéaire. Nous présenterons les formalismes par niveaux d'abstraction croissants.

De nombreux signes impliquent des primitives de mouvement simple des mains qui sont aisément représentables à l'aide de flèches. Il suffit alors pour représenter les signes, de fournir une ou plusieurs images du signeur permettant d'identifier les instants clés du signe (équivalent des tenues, contacts, configurations stables) et d'ajouter des symboles en sur-impression sur l'image pour indiquer le type de déplacement et les mouvements secondaires éventuels des mains. Une des application de notre travail de recherche a été de faciliter l'édition de tels *photosignes* par traitement automatique de vidéo.

Un dessin des contours des signeurs permet de mettre mieux en évidence les paramètres essentiels à la réalisation du signe. Ces représentations sont à la base de nombreux dictionnaires Français / LSF iconiques comme [MVGD97a] [Com06].

En représentant chacun des paramètres du mouvement de manière schématique, il est possible d'aboutir

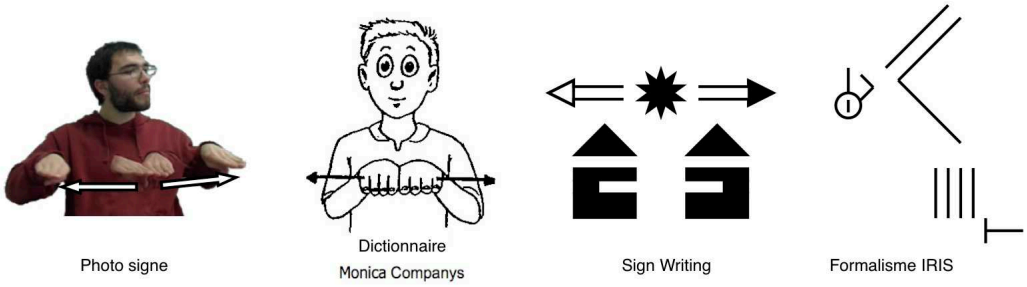


FIGURE 3.10 – Signe [TABLE] représenté à l'aide de différents formalismes 2D

à des formalismes permettant de rendre compte très précisément de la réalisation d'un signe et conservant la spatialisation des langues des signes. Le formalisme proposé par l'association IRIS[16] est un bon exemple de ces méthodes de description phonétique des Langue des Signes [GBM+07]. Il est actuellement en cours d'informatisation.

Le formalisme SignWriting [Sut08] créé par V. Sutton par adaptation d'un formalisme créé par la même auteur pour représenter les mouvements de danse, utilise également l'espace 2D pour retranscrire la spatialité du signe. Contrairement au formalisme IRIS, le plan image n'est pas forcément une projection simple de l'espace de signation. Les symboles conventionnels indiquant l'expression du visage, la configuration (incluant l'information d'orientation), et le mouvement sont disposés dans des vignettes de manière à représenter au mieux la structure réelle du signe, mais l'interprétation des vignettes est parfois subjective. En outre, ses promoteurs revendiquent la vocation de SignWriting à constituer un système d'écriture des LS en plus d'un système de notation.

3.2.7 Observations

L'utilisation des différentes méthodes de notation graphique ou informatique des signes a permis d'étudier plus en détail leur structure et de mettre en évidence un certain nombre de règles ou de tendances dans leur constitution. Ces règles semblent être peu dépendantes des langues des signes. Ces régularités sont-elles la conséquence d'une simplification des signes allant vers une facilité de production des signes ou d'une phonologie de la langue des signes ? Nous ne pouvons conclure sur cette question d'autant que le débat a toujours lieu au sein de la communauté de linguistes [Bou09]. Nous emploierons cependant le mot "phonologie" dans notre travail pour éviter les périphrases comme "régularité dans la structure des signes". Nous prenons le parti de nous attarder sur ces règles de phonologie car elles permettent d'ajouter des contraintes aux modèles de signes qui peuvent être utilisés par la suite dans le cadre du traitement automatique de productions en Langue des Signes pour accélérer ou rendre possible l'analyse de vidéos. Nous ne rentrons volontairement pas dans le détail des observations sur les configurations manuelles et les paramètres

[16] Institut de Recherche sur les Implications de la langue des Signes

non-manuels car notre travail se focalise sur le mouvement.

3.2.7.1 Observations synchroniques

Le fait de noter précisément les différentes étapes du mouvement permettent de mettre en évidence des phases de mouvement [KvGvdH98]. Les signes isolés sont précédés par une phase de *préparation* et suivis par une phase de *retrait*. Lors d'une utilisation des signes dans des énoncés, ces phases de préparation et de retrait ont tendance à être intégrées à des mouvements de *transition* permettant de passer d'une position finale d'un signe à la position initiale du signe suivant.

Battison a mis en évidence un certain nombre de règles communes à la plupart des signes de l'ASL. Nous avons pu vérifier que ces règles s'appliquaient également aux signes de la LSF. L'auteur distingue dans [Bat74] trois types de signes :

- Les deux mains bougent sans qu'une main dominante puisse être identifiée.
- Une seule main bouge mais les deux mains ont la même configuration manuelle.
- Une seule main bouge, mais les configurations manuelles sont différentes.

Battison note également un certain nombre de contraintes sur les points de contacts qui ne semblent pas pouvoir être combinées de n'importe quelle manière.

Il est également possible de remarquer que les différents mouvements impliqués dans la réalisation de signes appartiennent à un nombre restreint de primitives. Ainsi, [Uye97] montre que les plus fréquents sont les mouvements circulaires, les mouvements balistiques, les mouvements en "L", les mouvements en "X" et le mouvement en " ?"[17].

Au delà de ces considérations sur la cinématique du signe, il est intéressant d'examiner la dynamique des mouvements. Plusieurs linguistes ont formulé des observations intéressantes. Parmi eux, [Uye97] souligne le fait que "la durée d'un mouvement à répétition est de manière impressionnante, la même que celle qu'il faut pour articuler un mouvement simple". [KvGvdH98] fait également remarquer que les mouvements simples (sans répétition) sont beaucoup plus fréquemment suivis d'un temps de tenue que les signes répétés.

L'étude de la dynamique des signes donne lieu à une somme de travaux sur la *sonorité* des Langues des signes souvent basée sur l'analogie avec la succession de voyelles et de consonnes des langues vocales. Nous ne rentrerons pas dans le détail de ces travaux décrits dans [Bou09] bien que leur prise en compte constitue un indice intéressant dans le cadre du traitement automatique des langues des signes. Ceci constitue une perspective d'approfondissement de notre travail.

[17] Ces différentes primitives sont illustrées figure 7.3

3.2.7.2 Observations diachroniques

Si on s'attache maintenant à observer la variation de la structure des signes dans le temps, on note une simplification du mouvement tendant vers une économie articulatoire [L.00] ainsi qu'une économie de mouvement [GVKVDH98]. Cette économie dans la production pourrait être à l'origine des symétries qu'on observe également dans les gestes coverbaux [GVKVDH98]. Les *signes composés* résultant d'une mise bout à bout de deux signes standards tendent à se transformer de manière à ce que le signe résultant soit le plus régulier possible [Bat74]. On note toutefois une conservation des contacts [LJ86].
Cependant, les signes évoluent également de manière à conserver leur potentiel iconique, ce qui va parfois à l'encontre des règles phonologiques. C'est ce qui permet le va et viens continuel entre les visées illustrative et non-illustrative.

3.2.8 Bilan

Pour conclure cette partie sur le signe et les manières de le transcrire graphiquement et informatiquement, nous ne pouvons que souligner la grande diversité des formalismes proposés qui permettent, tous à leur manière, d'apporter un éclairage sur ce qu'est le signe. On peut distinguer deux types d'approches :

Les approches multisegmentales permettent de noter phonétiquement la manière dont sont réalisés les signes avec peu de prises de parti apparentes. Elles permettent d'étayer des hypothèses sur la structure des signes. L'inconvénient de ces approches est la grande sur-spécification des signes. Ces formalismes permettent d'écrire des signes absolument incohérents phonologiquement (ce qui n'implique pas pour autant que ces méthodes de notation soient incorrectes).

Les approches monosegmentales décrivent les signes de manière plus synthétiques. Les hypothèses sous-jacente sur la nature des signes sont souvent très restrictives si bien que certains des formalismes sont mal adaptés pour décrire les signes composés et les unités gestuelles de grande iconicité.

En pratique, on observe que les méthodes de transcription tendent à converger :

- Dans les signes à répétitions, même les approches multisegmentales décrivent les signes en utilisant des raccourcis de notation évoquant les répétitions (alors même que les deux occurrences répétées ne sont pas identiques sur la vidéo).

- Lorsque les signes deviennent trop compliqués, les approches monosegmentales les considèrent comme une succession de signes élémentaires).

Nous utiliserons dans notre travail de recherche un modèle inspiré des approches monosegmentales car les descriptions sont plus synthétiques et rendent mieux compte de la variabilité des signes.

Le fait de noter systématiquement la structure des signes permet de faire ressortir des règles de structure respectées par la plupart des signes standards. Nous utiliserons dans notre travail de recherche les principes suivants :

- L'importance des phases de préparation et de retrait ainsi que la dynamique du signe [KvGvdH98]
- La possibilité de décrire la plupart des signes par un ensemble restreint de primitives [Uye97]
- Les règles concernant la réalisation d'un signe à une main à l'aide de la main dominante [Bat74]
- Les règles de symétrie des mouvements des deux mains lorsqu'un signe implique les deux mains [Bat74]
- L'importance des contacts dans le signe (donc de la position relative des deux mains) [Bat74]
- La pertinence de la notion de répétition dans la définition du signe [Cha02]
- La régularité des signes et le fait que les signes non répétés sont plus souvent suivis d'une phase de tenue [KvGvdH98]
- La variabilité signifiante des signes affectant la position, le mouvement et l'orientation des signes [Cux00]

Toutes les règles que nous venons d'énoncer peuvent être transgressées dans certains signes car tout signe est le produit d'un certain nombre de contraintes (fig. 3.11) dont certaines vont à l'encontre de la phonologie. Pour cette raison, il est nécessaire d'envisager la notion de signe sous une approche morphophonétique [Bou07] permettant de prendre simultanément en compte les plans de l'expression et du contenu.

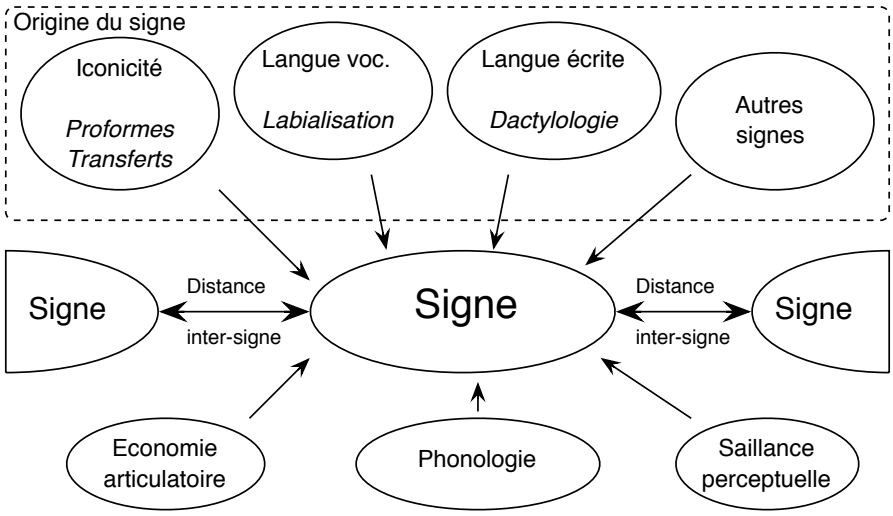


FIGURE 3.11 – Bilan des forces agissant dans la stabilisation des signes standards

Deuxième partie

État de l’art sur le traitement automatique des Langues des Signes

CHAPITRE 4

DIFFÉRENTES MÉTHODES D'ACQUISITION D'UNE PRODUCTION EN LANGUE DES SIGNES

Cette deuxième partie de l'état de l'art est consacrée aux techniques d'acquisition d'une production en Langue des Signes. Nous passerons en revue dans un premier temps, différentes méthodes de capture de mouvement qui sont couramment utilisées pour l'étude du mouvement communicatif, puis nous expliquerons les motivations qui nous ont poussé à utiliser les données vidéos mono-vue dans le cadre de notre travail de recherche.
Si les techniques de reconstruction de la postures basées sur une simple vidéo sont intéressantes du point de vue des nombreuses applications qu'elles rendent possibles, elles doivent également résoudre différents problèmes que nous détaillerons.
Nous proposerons une analyse des différentes méthodes mises en oeuvre pour effectuer le suivi et la reconstruction 3D d'une posture de la partie supérieure du corps à partir de vidéos mono-vue en mettant en avant leur adéquation avec les spécificités des Langues des Signes.

4.1 Différents systèmes d'acquisition du mouvement

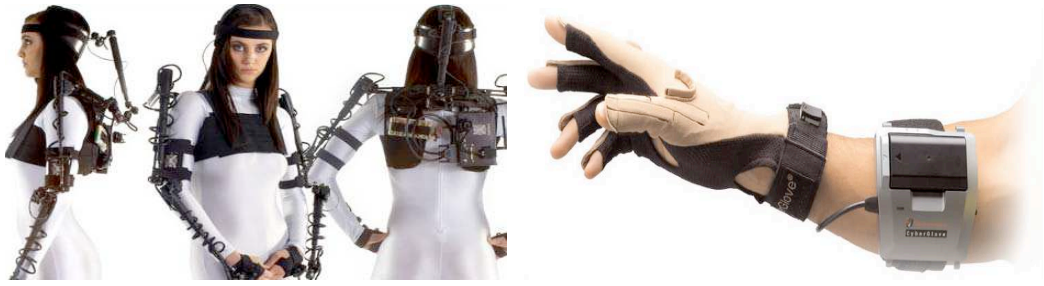
Le but des systèmes de capture de mouvement (*mocap*) est de fournir à chaque instant une description de la configuration spatiale d'un système. Dans notre cas, le système dont il est question est la partie supérieure du corps humain.
Différents critères peuvent orienter vers le choix d'un dispositif de capture de mouvement. Parmi eux, nous pouvons citer la précision des mesures, la sensibilité des appareils de mesure aux conditions d'acquisition, les contraintes sur les mouvements du sujet, les contraintes vestimentaires du sujet, le coût des dispositifs [1], la complexité d'utilisation des appareils utilisés pour l'acquisition, le caractère plus ou moins invasif des dispositifs et la nécessité de calibrage.
Nous détaillons dans cette partie les spécificités des différents dispositifs de capture de mouvement en mettant en avant leur adéquation avec l'étude des Langues des Signes dans un but d'analyse ou de génération.

4.1.1 Dispositifs electro-mécaniques

Les dispositifs de capture de mouvements mécaniques sont des exo-squelettes (fig. 4.1).

La posture est estimée grâce à des mesures d'angles aux différentes articulations (épaule, coude, dos , nuque ...) par le biais de dispositifs electro-mécaniques. Le système a l'inconvénient de restreindre la liberté de mouvement du signeur. Les mesures acquises permettent de reconstruire directement la posture du signeur sans avoir à recourir à une étape de cinématique inverse. La précision du système est limitée en ce

[1] L'ensemble des prix et des caractéristiques techniques que nous présentons datent de janvier 2010.



Exo-squelette (www.metamotion.com) Cyberglove (www.inition.co.uk)

FIGURE 4.1 – Dispositifs de capture de mouvement électro-mécaniques

qui concerne la restitution de contacts particulièrement importants en Langue des Signes. Ceci est aisément compréhensible car on cumule les erreurs de mesures d'angles sur l'ensemble de la chaîne cinématique reliant les deux points en contact (Actuellement, l'erreur est environ de 0.5 degré pour chaque angle mesuré dans les gants[2]). Les dispositifs d'acquisition sont relativement coûteux. A titre indicatif, le seul exo-squelette pour le haut du corps coûte 12 k$[3]. Il est à noter que des gants comme le système cyberglove présenté (fig. 4.1) permettent également de mesurer la configuration manuelle. Ce sytème est entre autre utilisé par [MGWW00] pour réaliser de la capture de mouvement en vue de traduction automatique de la LS chinoise.

4.1.2 Dispositifs magnétiques

Les dispositifs d'acquisition magnétique du mouvement utilisent un émetteur de champ magnétique ainsi que des capteurs magnétiques répartis sur les différents membres du sujet dont on souhaite effectuer le suivi. Ils sont utilisables pour effectuer un suivi de la tête, du buste, des épaules, des bras et des avants bras du signeur. L'encombrement des capteurs (fig. 4.2) rend toutefois impossible l'utilisation de ce dispositif pour effectuer une reconstruction de la configuration manuelle (qui est alors souvent effectuée à l'aide de la méthode électro-mécanique).

La précision annoncée par les concepteurs est, pour un éloignement d'1m50 du bloc émetteur, de l'ordre du cm pour les positions et du degré pour les orientations fournies par les capteurs. Le champ peut être perturbé par des objets métalliques qui diminuent la précision des données obtenues ou engendrent un bruit de mesure qu'il faut ensuite éliminer par filtrage. Nous avons eu l'occasion d'utiliser ce dispositif pour capturer des données permettant de valider nos techniques de reconstruction de la posture et avons été confronté à ces nombreux artefacts. La reconstruction de posture met en oeuvre des techniques de cinématiques inverse. Le coût d'achat des équipements de capture de mouvement est de l'ordre de 30k$[4].

[2]http ://www.vrealities.com/cyber.html
[3]http ://www.metamotion.com/motion-capture/electro-mechanical-motion-capture.htm
[4]http ://www.ascension-tech.com/realtime/MotionSTARWirelessLITE.php

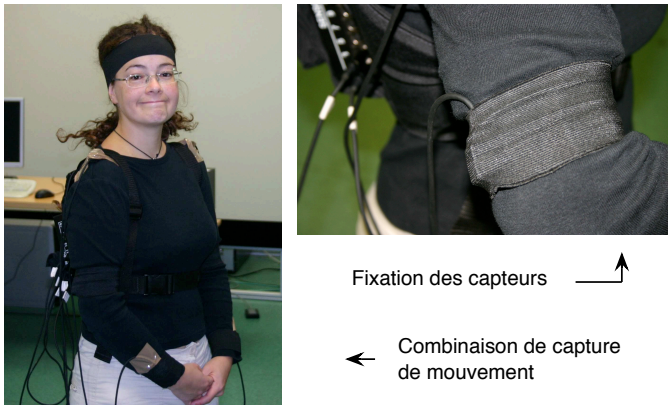


FIGURE 4.2 – Capture de mouvement à l'aide d'un dispositif magnétique

4.1.3 Dispositifs de capture intertiels

Dans le cadre de la capture inertielle, les capteurs situés sur les articulations permettent de mesurer une orientation absolue du capteur grâce à des gyroscopes. Chaque capteur fournit aussi une estimation des accélérations qu'il subit. L'encombrement des capteurs est relativement important, même si leur miniaturisation a rendu récemment possible la création de gants dont certains ont été utilisés pour effectuer de la reconnaissance de dactylologie[5] [HREAU08]. La précision annoncée par les constructeurs est de l'ordre du degré[6] mais de la même manière que pour les dispositifs mécaniques, on cumule les incertitudes sur la chaîne cinématique. Un système de capture complet du haut du corps coûte approximativement 2k$ [Kub07].

Dispositif de capture inertiel
(www.xsens.com)

Jeu AcceleSpell (TM) développée à partir du
gant de capture AcceleGlove (TM)

FIGURE 4.3 – Capture inertielle de la posture et des configurations manuelles

[5]http ://www.acceleglove.com/applications.asp
[6]http ://www.xsens.com/images/stories/products/PDF Brochures

4.1.4 Dispositifs optique avec marqueurs

Une technique de capture de mouvement assez répandue pour sa grande précision est la capture de mouvement optique où le corps du signeur est couvert de petits marqueurs. On distingue les techniques basées sur les marqueurs actifs et sur les marqueurs passifs.

4.1.4.1 Marqueurs actifs

Dans cette technique, les marqueurs sont des micro-leds émettant des signaux captés par des cellules photosensibles situées dans des bases d'acquisition. Chaque marqueur peut être identifié de manière unique grâce au signal qu'il émet. L'inconvénient de ce système est l'encombrement important des marqueurs (20 x 14 x 3.2 mm). Des gants basés sur ce système permettent la capture des configurations manuelles. Il est également nécessaire d'avoir plusieurs bases d'acquisition pour permettre une visibilité satisfaisante des différents capteurs malgré les occultations partielles. L'erreur moyenne de ce type de dispositif est de 0,3 mm [Vie00].

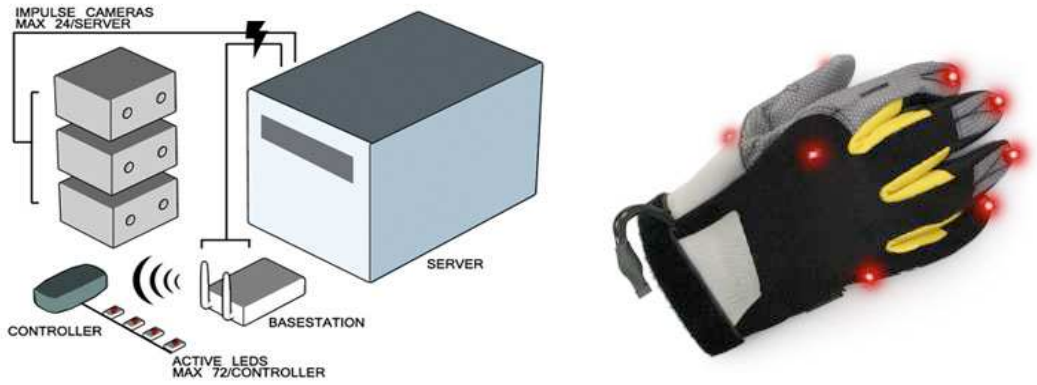


FIGURE 4.4 – Illustration du fonctionnement de la capture de mouvement vidéo par marqueurs actifs et du gant pour la capture des configurations manuelles (http ://www.phasespace.com)

4.1.4.2 Marqueurs passifs

La capture de mouvements basée sur les marqueurs passifs utilise des marqueurs réfléchissants plats ou sphériques disposés à la surface des articulations du signeur. Leur faible taille (de 3 à 10 mm de diamètre) permet de suivre avec la même technique les déformations du visage, la posture du signeur et la configuration manuelle. Les mouvements sont filmés à l'aide d'un nombre de caméras infra-rouge qui varie en fonction de la complexité du mouvement. Les différents points de vue permettent de limiter les occultations totales des marqueurs et de faciliter leur appariement (l'une des difficultés de cette technique est en effet d'associer sur chaque vue un point lumineux à un marqueur précis). L'utilisation de multiples caméras

implique une étape de calibration relativement longue. Cette technique permet d'estimer la position des marqueurs avec une erreur moyenne de 0,3 mm [Vie00]. L'inconvénient notoire de cette technique est son coût d'acquisition de 150 k$ à 250 k$[7]. Les essais que nous avons eu l'occasion de mener avec cette technique ont prouvé qu'elle était satisfaisante pour la capture et la synthèse par avatar des postures et des configurations manuelles du signeur.

FIGURE 4.5 – Capture de mouvement basée sur les marqueurs passifs

4.1.5 Dispositifs optique sans marqueurs

L'acquisition optique de postures du corps humain sans l'usage de marqueur est encore une thématique de recherche active même si des applications commerciales commencent à voir le jour (on pourra citer à titre d'exemple Organic Motion Stage : développé par la société Organic Motion[8]). L'intérêt de toutes ces méthodes est leur coût moindre comparativement aux dispositifs que nous venons de citer précédemment.

4.1.5.1 Utilisation de multiples caméras

Il existe plusieurs méthodes de reconstruction de la posture à partir de caméra multiples.

La première famille de méthodes consiste à utiliser un grand nombre de caméras afin d'obtenir pour chacune une silhouette du personnage. Le croisement des différentes silhouettes permet d'obtenir une enveloppe du corps humain sous la forme d'un polyèdre (ou ensemble de voxels). La méthode utilise un modèle volumique du personnage sous forme de primitives géométriques. Il s'agit ensuite de retrouver pour chaque image la posture du modèle, de telle manière à ce que le modèle soit le plus possible contenu dans l'enveloppe et tangent à chaque face du polyèdre. Une telle méthode est décrite dans [BMGB09]. Les méthodes commerciales utilisant cette approche reportent une erreur de 1 à 2° sur l'estimation des angles des articulations[9]. Lorsque les caméras sont moins nombreuses, le but des méthodes consiste à recaler le modèle 3D sur les silhouettes de chaque vue [BSB05] [Nor07].

[7] http ://www.metamotion.com/motion-capture/optical-motion-capture-1.htm
[8] http ://www.organicmotion.com/
[9] http ://www.organicmotion.com/

La deuxième famille de méthodes fait appel à moins de caméras et utilise plus d'indices visuels sur le personnage à suivre (par exemple la texture, la couleur de peau, la forme des membres). L'utilisation de multiples vues permet de mieux traiter les occultations en utilisant de préférence la ou les vues ou les objets à suivre ne seront pas occultés. On citera à titre d'exemple [Fon08].

La troisième famille de méthode utilise l'appariement stéréoscopique de plusieurs vues prises de points relativement proches pour reconstruire une image avec profondeur (on notera que de telles cartes de profondeur[10] peuvent également s'obtenir à l'aide de caméra temps de vol). Le traitement de vidéos avec carte de profondeur présente également de nombreux avantages dans la mesure où il est plus facile de segmenter le personnage du fond et de distinguer des objets ayant une forme similaire (par exemple, il est possible de segmenter la forme de la main, même quand celle-ci occulte la tête). De telles travaux de recherche sont décrits dans [KVD06].

4.1.5.2 Utilisation d'une seule caméra

Nous ne développerons pas dans ce paragraphe les différentes méthodes consacrées au suivi de la posture à l'aide d'une seule vue car ce problème fait l'objet du paragraphe suivant. Nous soulignerons juste le fait qu'à l'opposé des approches que nous avons décrites jusqu'ici, une simple vue ne contient *a priori* pas assez d'information pour reconstruire précisément la posture d'un signeur. Il s'agit donc d'un problème mal posé beaucoup plus difficile à résoudre.

4.1.6 Synthèse

La partie qui précède a permis de faire un tour d'horizon des techniques existantes pour reconstruire la posture d'un signeur. Plusieurs points ressortent de cette comparaison :

- Les dispositifs spécialement conçus pour la capture de mouvement sont relativement coûteux (au moins 2000$) et donc difficilement accessibles au grand public. Il est cependant envisageable de les utiliser pour élaborer des modèles de Langue des Signes ou pour capturer des mouvements en vue de synthèse de LS par des signeurs virtuels.

- Les dispositifs utilisant des capteurs trop invasifs comme les gants de capture limitent les mouvements du signeur et nous semblent donc mal adaptés à l'étude des langues des signes, même si certaines équipes de recherches sur les LS les utilisent pour acquérir des données [MGWW00] [Vog03].

- Les dispositifs utilisant des caméras multiples sont relativement contraignants car ils nécessitent une étape de calibration et un aménagement de l'espace de capture de mouvement. Ils ne sont donc pas facilement utilisables dans le cadre d'une application grand public.

[10] Image dans laquelle chaque pixel contient une valeur de profondeur par rapport à la position d'observation.

- Le fait que les dispositifs d'acquisition soient très visibles dans l'environnement du signeur peut aussi avoir une influence sur sa manière de signer. Il y a alors un risque que la langue des signes produite soit moins naturelle.
- Les dispositifs d'acquisition stéréo ou basés sur des caméras temps de vol sont encore difficilement accessibles pour une application grand public mais on peut émettre l'hypothèse que les nouvelles générations d'ordinateurs les intégreront peut être à moyen terme. Il serait alors possible d'imaginer des utilisations de vidéos intégrant la profondeur à partir d'ordinateurs personnels. Par contre, le fait de baser une technique de reconstruction de la posture uniquement sur ces technologies interdit le traitement de vidéos mono-vue dont nous disposons actuellement (qui sont pour leur immense majorité des vidéos 2D).

Pour les raisons qui précèdent, nous avons donc choisi de nous concentrer sur le développement de techniques de traitement et d'analyse de vidéos mono-vue, même si notre travail a ponctuellement fait appel aux techniques de capture de mouvement traditionnelles (capture de mouvements magnétique et optique avec marqueurs passifs) pour valider des modèles ou des méthodes.

4.2 Reconstruction monoculaire de la posture

Si le terme *monoculaire* fait explicitement référence au fait que la séquence est filmée à l'aide d'une seule caméra, il est nécessaire à ce point de notre étude de définir précisément la notion de *posture* telle que nous l'entendons dans cet ouvrage. En langue des signes, le problème de capture d'énoncés peut être présenté comme la somme de trois-sous problèmes :

Le premier est celui de la reconstruction de la posture de la partie supérieure du corps (en excluant toutefois la configuration et l'orientation et la main qui sont beaucoup plus difficile à déterminer à partir de la vidéo). Une posture sera donc caractérisée dans notre travail, par un ensemble de mesures d'angle ou de position permettant de spécifier les positions relatives des mains, des avants-bras, des bras, du buste, du cou et de la tête.

Le second concerne le suivi ou l'identification des configurations manuelles. Il s'agit d'un domaine à part, en raison des difficultés à surmonter : nombre important d'auto-occultations, flou, variations extrêmement rapides des configurations manuelles. En terme de suivi, même les études récentes comme [DM06] utilisant plusieurs vues atteignent des résultats difficilement utilisables dans le cadre des LS (Ils utilisent l'hypothèse de configurations invariantes durant l'ensemble du signe et sélectionnent manuellement les points caractéristiques de la main du signeur). En terme de reconnaissance à partir d'une seule vue, les travaux se limitent à une dizaine de configurations identifiées et interdisent souvent les rotations hors plan [MB99] [TvdM01] [TPNY02] [YCW$^+$08]. Ces hypothèses sont naturellement en contradiction avec la nature de la LSF qui implique des rotations hors plan des mains et compte plus de 50 configurations.

Le troisième sous-problème porte sur la capture des expressions du visage. Ce problème se trouve à mi chemin entre la 2D et la 3D et peut être traité comme l'estimation d'une déformation de texture. Actuellement, les approches monovues commencent à suivre ces déformations, même en présence d'occultations du visage [Mer07] même si la plupart d'entre elles nécessitent un apprentissage de l'apparence du signeur et une vue de face de sa tête.

Dans cette section, nous nous intéressons aux différentes techniques qui ont déjà été mises en oeuvre pour reconstruire la posture 3D d'une personne à partir d'une vidéo monovue. Une vision générale du domaine nous est proposée par [MG01] ; [Fon08] présente également un intéressant tableau comparatif des différentes méthodes de reconstruction de la posture, en mentionnant les modèles de corps humain utilisés et le protocole d'acquisition des données traitées (nombre de caméras, marqueurs additionnels).
Dans un premier temps, nous nous intéresserons aux spécificités du problème de traitement de la vidéos en LS dans un but de reconstruction de la posture. Cette étape d'analyse des hypothèses de travail est d'autant plus importante que la plupart des techniques de reconstructions de la posture ont été conçues pour des domaines spécifiques (interaction homme-machine, réalité virtuelle, jeux vidéos ...) et ont donc des contraintes bien différentes.
Dans un second temps, nous détaillerons les grandes étapes nécessaires à la reconstruction de la posture 3D d'un signeur en illustrant chaque étape à l'aide de travaux de recherche. Les méthodes décrites dans la littérature étant nombreuses, nous mentionnerons en priorité les travaux portant sur des vidéos monovue, visant à reconstruire la posture de la partie supérieure du corps et ne faisant pas usage de marqueurs additionnels.

4.2.1 Type de vidéo traité

Le problème de reconstruction de la posture à partir d'une seule vidéo pose de nombreux problèmes. Il est donc nécessaire d'ajouter des hypothèses simplificatrices. On peut aussi ajouter des données (autres qu'une deuxième vue) ou des connaissances (modèle du corps humain …) supplémentaires acquises par apprentissage et correspondant entre autre aux hypothèses marquées *. Nous les précisons au §4.2.5.
[MG01] classe ces hypothèses entre les hypothèses sur le mouvement et les hypothèses sur l'apparence. Cette classification se trouve dans le tableau 4.6.

Dans notre cas, les contraintes ont été choisies de manière à correspondre à des conditions de compréhensibilité optimale de LS. Une description de cet environnement est proposée dans l'ouvrage [BEJ07] portant sur les spécificités de l'interprétation en LS.

Contraintes sur l'environnement : Il est préférable que le signeur se positionne devant un fond clair et évite les positions de contre jour tout en restant correctement éclairé.

Contraintes sur l'apparence du signeur : Le signeur doit porter un haut uni, si possible sombre, évitant les objets brillants qui perturberaient la vision.

Hypothèses sur le mouvement	Hypothèses sur l'apparence
Sujet restant dans l'espace de travail	**Hypothèses sur l'environnement**
Caméra statique ou avec un déplacement connu	
Une seule personne dans l'espace de travail	Eclairage constant
Sujet constamment filmé de face	Fond fixe
Mouvement parallèle au plan de la caméra	Fond uniforme
Pas d'occultations	Paramètres de la caméra connus*
Mouvements lents ou continus	Appareillage spécial
Membres limités en mouvement	
Type de mouvement du sujet connu d'avance	**Hypothèses sur le sujet**
Sujet évoluant sur un plan horizontal	
	Pose de départ connue*
	Sujet connu*
	Présence de marqueurs sur les sujet
	Vêtements colorés spéciaux*
	Vêtements proches du corps

FIGURE 4.6 – Hypothèses les plus fréquentes pour la capture de mouvements mono-vue extraites de [MG01]

Contrainte de posture : Il est préférable de faire face à son interlocuteur pour une compréhension optimale.

Contrainte de cadrage [11] : Le cadrage doit être suffisant pour permettre de voir au moins la tête et les mains du signeur lorsqu'elles contribuent à transmettre le message.

Ces conditions sont certes restrictives, mais correspondent à des vidéos qu'il est possible de trouver sur des sites comme celui de Websourd[12]. On remarquera que les contraintes sur l'apparence du signeur sont en contradictions avec l'utilisation de marqueurs colorés comme ceux utilisés par [HGO96] ou [YC00]. Nous avons donc pris le parti de ne pas en utiliser. En revanche, il est important de souligner qu'aucune restriction n'est faite sur la LSF qui se situe au coeur de notre étude.

4.2.2 Hypothèses simplificatrices pour nos vidéo en LSF

Les bonnes pratiques que nous venons de citer permettent de formuler un certain nombre d'hypothèses que nous pourrons réutiliser dans le cadre du traitement automatique de vidéos. Nous ajouterons le signe (*) à côté de toutes les hypothèses qui découlent directement du § 4.2.1.

Contrainte sur le fond* : Le fond est fixe et clair, pas forcément uniforme, d'une couleur bien différente de celle du signeur. Il sera donc possible d'obtenir facilement la silhouette du signeur.

[11] il s'agit là d'une contrainte que nous ajoutons, propre à la capture vidéo, non mentionnée dans l'ouvrage [BEJ07]
[12] www.websourd.org

Contrainte sur l'apparence du signeur* : Le signeur porte un haut uni et sombre, donc différent du fond et de la couleur de peau.

Contrainte sur le contenu de la vidéo : La vidéo contient un seul signeur et un fond.

Contrainte sur la morphologie du signeur : La morphologie du signeur ne varie pas durant toute la vidéo.

Contrainte sur l'illumination* : L'éclairage est constant, donc les couleurs des différentes parties du corps varient peu[13].

Contraintes sur le cadrage* : Le cadrage est effectué de manière à ce que le signeur soit vu de face, donc l'apparence du buste est peu variable[14].

4.2.3 Difficultés propres aux vidéos en Langue des Signes

Même si les hypothèses que nous venons de citer permettent de simplifier grandement le traitement des vidéos, il est important de souligner des problèmes présents dans certains autres domaines de la capture de mouvements qui prennent une grande ampleur dans le cas des LS.

- Les mouvements sont rapides et de grande amplitude. Le plus souvent, les vidéos que nous devons traiter ont une fréquence d'acquisition de 20 à 30 images par seconde. En conséquence, il n'est pas possible d'utiliser d'hypothèses de petits déplacements. Ceci rend notamment difficile l'application de la plupart des méthodes basées sur les flots optiques.

- Une autre conséquence de la vitesse des mouvements est que les images des mains obtenues sont souvent légèrement floues, ce qui complique beaucoup l'identification des configurations impliquées dans les signes. Ce phénomène de flou est souvent accentué par la compression des vidéos.

- Les mouvements sont non-linéaires et contiennent de nombreux points de rebroussements qui ont une importance dans la transmission du message. Il est donc difficile de prédire le mouvement avec des approches traditionnelles comme les prédicteurs de Kalman. De plus, il faut être vigilant avec l'emploi de méthodes de suppression de bruit de mesure qui ont tendances à lisser les trajectoires aux points de rebroussements.

- Certaines parties du corps à suivre ont une apparence relativement proche (par exemple, les deux mains ont la même couleur et des formes et textures voisines). Il est nécessaire de gérer aussi bien les occultations que les sorties de cadre de ces parties du corps. A tout moment, il faut que le système puisse permettre d'assigner chaque pixel de l'image à un membre du signeur, d'où certains problèmes de désambigüisation (comme distinguer la main droite de la main gauche).

[13] Il y a tout de même de légères variations de luminosité dues à la variation de l'orientation des surfaces.
[14] Il est toutefois possible de modifier légèrement l'orientation du buste par rapport à la caméra lors de transferts personnels.

- Les occultations induisent une variation importante de l'apparence de certains membres comme les épaules et les coudes.
- L'apparence du signeur est susceptible de varier d'une vidéo à l'autre. La variation portera à la fois sur la couleur de peau du signeur, sa morphologie et les habits qu'il porte.
- Même si l'apparence du signeur est contrainte (vêtements unis), il faut être en mesure de traiter les vidéos où le signeur ne porte pas nécessairement des manches longues.

4.2.4 Méthodes génériques

Après avoir pointé clairement les hypothèses sur les vidéos que nous allons traiter, ainsi que les problèmes à surmonter lors de l'élaboration de systèmes de reconstruction de la posture 3D d'un signeur à partir d'une vidéo monoculaire, cette partie expose les travaux déjà effectués dans le domaine.

Différentes méthodes génériques existent pour reconstruire la configuration géométrique d'un système à partir d'une vidéo. Nous montrons pourquoi aucune d'elles n'est vraiment facile à utiliser dans le cas qui nous intéresse.
Les méthodes **Shape from shading** [PF04] [DFS08] permettent de retrouver la profondeur d'une image en utilisant la brillance. Les hypothèses sur lesquelles sont fondées ces méthodes (surfaces lambertiennes, sources de lumière clairement caractérisées) ne sont pas vérifiées dans notre cas.
Les méthodes **Shape from texture** [For01] ont pour fonction de retrouver la géométrie d'un modèle en utilisant la projection locale de sa texture. Elles nécessitent donc des solides fortement texturés (ce qui est contraire à nos hypothèses de départ).
Les méthodes **Shape from focus/défocus** [Sta76] permettent de retrouver la profondeur des points d'un système en utilisant les réglages de focale permettant de le voir nettement dans l'image. Ceci nécessiterait de capturer la même image avec plusieurs focales différentes.
Les méthodes **Shape from motion** [TK92] utilisent les informations de flot optique pour reconstruire la géométrie d'un système. Elles sont extrêmement délicates à mettre en oeuvre dans le cadre de la LSF en raison des occultations multiples, de la rapidité d'exécution des signes[15], du faible texturage, et du fait que le corps humain est articulé et non rigide. [DM06] utilise cette approche pour reconstruire les configurations manuelles mais l'auteur pointe manuellement les points caractéristiques des doigts sur chaque vue et prend le soin d'ajouter un bruit gaussien pour prouver la robustesse de son approche ...

4.2.5 Comparaison des méthodes de reconstruction mono-vue de la posture

D'autres méthodes sont applicables, car on connait un modèle préalable du système à suivre (§ 4.2.2).

[15]Ces mouvements rapides vont à l'encontre d'hypothèses de petits mouvements souvent utilisée dans les méthodes de résolutions d'équations de shape from motion

Modèle d'apparence : Modèles de couleur (peau, fond ou vêtements) , apparence des différents membres (contours, formes),

Modèle de géométrie du corps humain : Morphologie du signeur, degrés de mobilité, contraintes de non-collision,

Modèle de posture : Butées articulaires ou distribution des postures impliquées dans la production de signes (certaines postures sont plus confortables ou plus fréquentes que d'autres),

Modèle de mouvement : Continuité du mouvement,

Modèle de signe : Utilisation de contraintes phonologiques (modèle explicite) ou exploitation d'une base de donnée (modèle implicite) de signes appris qui indiquent les signes vraisemblables en LSF,

Modèle de LSF : Connaissances de plus haut niveau sur la LSF (syntaxe, prosodie).

Les méthodes de la littérature consacrées aux gestes que nous allons exposer tirent parti de ces différents modèles à l'exception des deux derniers directement en lien avec la LSF. Nous présenterons dans le §6 comment la prise en compte des spécificités de la LSF dès le traitement bas niveau des vidéos en LS permet d'améliorer la reconstruction de la posture.

Les différentes méthodes dédiées à une reconstruction mono-vue de la posture suivent en général le schéma suivant :

- Initialisation du suivi (apprentissage de l'apparence du signeur et du fond, initialisation de la morphologie du signeur, contrainte de la posture de départ),
- Prétraitement de la vidéo (changement d'espace de représentation, détections de formes géométriques génériques),
- Détection du signeur,
- Détection locale de parties du corps ou d'autres indices locaux,
- Reconstruction de la posture 2D puis 3D du signeur,

Il est important de souligner que toutes ces étapes ne sont pas forcément nécessaires à chaque technique de reconstruction 3D de la posture et que certaines peuvent être imbriquées. Ainsi, certaines méthodes ne nécessitent que peu d'initalisation, n'utilisent pas d'indices locaux ou reconstruisent directement la posture 3D du signeur sans passer par une représentation 2D.

En raison du grand nombre de travaux qui concernent la reconstruction 3D de la posture, il ne nous sera pas possible de décrire en détail chacune des méthodes figurant dans la littérature. Les états de l'art existants proposent plusieurs classifications des différentes des méthodes. Certaines distinguent les approches 2D et

3D. D'autres proposent de classer les méthodes suivant que les approches utilisés sont *basées modèle* (le but est alors de trouver la posture qui minimise l'écart du modèle à l'image courante) ou qu'elles sont *basées image* (l'image courante est comparée à une base d'apprentissage pour lesquelles les postures correspondantes sont connues). Cette taxonomie nous semble de plus en plus délicate à mettre en oeuvre, car on observe que les méthodes les plus efficaces mêlent les approches basées 2D et 3D et ne sont pas forcément exclusivement basées modèle ou image. Dans cet état de l'art, nous proposons donc, à l'instar de [MG01], une comparaison des techniques basée sur les différents modèles utilisés et sur les étapes de traitement de la vidéo.

4.2.6 Prétraitements de la vidéo

Le premier prétraitement consiste à calculer une appartenance de chaque pixel à différentes classe à partir de sa couleur. Ceci permet entre autre de **détecter les pixels de peau**.
Dans un premier temps, il est important de choisir un espace de représentation de la couleur. Nous ne citerons dans les lignes qui suivent que les espaces de représentation les plus utilisés en traitement d'image :

- L'espace RGB (Rouge Vert Bleu) est utilisé pour représenter la couleur dans un but de synthèse additive. Il est utilisé par [ILI98] pour effectuer un suivi de la tête et des mains.

- L'espace RGB normalisé s'obtient en normant chacune des composantes RGB par la somme (R+G+B). On peut ainsi, en première approximation, considérer que toutes les couleurs ont la même luminosité. Le passage dans cet espace permet notamment d'être moins dépendant des ombres ou changements d'éclairement de la peau. Cet espace est utilisé par [GCG$^+$96] pour effectuer des détections de visages.

- Dans l'espace HSV (Teinte Saturation Valeur), la Saturation représente la pureté de la couleur (entre 0 : gris et 1 : couleur pure) et la Valeur correspond approximativement à la luminosité.

- Dans les espaces YCrCb et YUV[16], l'intensité Y est calculée en pondérant les composantes Rouge Vert et Bleu de manière à tenir compte de la différence de sensibilité de l'oeil aux différentes longueurs d'ondes. Ce type d'espace est entre autre utilisé par [YA98] et [HLM04] pour détecter les pixels de peau.

Les études menées par [TA00] ont montré que les résultats de détection de peau étaient meilleurs dans les espaces non normalisés (ce qui parait normal car on dispose alors de plus d'informations sur la couleur) et que la chrominance jouait un rôle plus important que la luminance. D'autres études comme [MSL01] ont également montré que la chrominance de la peau variait en fonction de son illumination lors de sur- ou de sous-éclairement. Ceci pourrait également expliquer l'importance de la prise en compte de la luminance.
Une fois l'espace de travail choisi, il est possible d'exploiter plusieurs types de modèles :

[16] D'autres espaces de couleurs comme les espaces Lab, pourraient être inclus dans cette catégorie. Nous ne les développons pas pour des raisons de place.

- Les premières familles de modèles indiquent l'**appartenance ou non d'un pixel à une classe** en se basant sur sa couleur. Il est possible de délimiter ces classes par des hyperplans dans l'espace de couleur ou d'utiliser des délimitations à l'aide de surfaces plus élaborées par l'intermédiaire de méthodes de classification non linéaires comme les SVM[17].

- Les secondes familles de modèles fournissent une **probabilité que le pixel appartienne à une classe** en se basant sur sa couleur. Les modèles mathématiques peuvent par exemple être spécifiés sous forme de gaussiennes ou de mélanges de gaussiennes. Il est aussi possible de spécifier des formes de distributions plus élaborées à l'aide de RVM[18]. L'utilisation d'histogrammes de couleur de peau et de couleur de l'image permet aussi d'estimer *a posteriori* la probabilité qu'une couleur appartienne à la catégorie peau en utilisant une approche bayésienne [Gia08].

Les différents modèles de couleur utilisés peuvent être fixés avant le déroulement de l'algorithme de reconstruction de la posture. Ceci est en particulier le cas pour les modèles de couleur de peau, pour lesquels il est assez facile de donner un modèle relativement générique. Il peut également être intéressant de fixer le modèle lors de l'initialisation des méthodes de suivi. On peut ainsi adapter le modèle à la couleur de la peau du signeur de manière à la discriminer au mieux du fond de la vidéo à traiter, ou de prendre en compte des couleurs de peau très différentes.
Enfin, il est possible d'adapter le modèle au cours du suivi comme le font [LW07] ou [WS06]. Un danger de ce type d'approche est que le modèle peut évoluer jusqu'à ne plus être cohérent.
De nombreuses études ont comparé la pertinence des différents espaces de couleurs pour la segmentation de la peau. Les auteurs de [BM00b] concluent que les approches bayésiennes de classification conduisent à moins de faux positifs pour la détection des pixels de peau. D'autre part, [TA00] montre qu'une simple distribution gaussienne dans les espaces des couleurs modélise bien la classe de couleur peau. Les modélisations plus élaborées comme les multigaussiennes n'apportent pas d'amélioration très significatives lorsqu'on se trouve dans une situation d'éclairage maîtrisé.

D'autres prétraitements de la vidéo peuvent permettre de **séparer la silhouette du signeur et le fond de la vidéo**. Une première approche pour y parvenir est d'utiliser l'appartenance de la couleur des pixels aux classes de couleur "fond" et "non-fond". Les techniques que nous venons d'évoquer dans la partie précédente sont alors applicables.
Si le fond est fixe et éclairé avec une illumination constante au cours de la vidéo, une simple soustraction entre l'image courante et l'image de fond, suivie d'un seuillage et d'opérations morphologiques peut permettre d'obtenir la silhouette du signeur. Cela nécessite naturellement de connaître l'image de fond ou

[17]Les SVM (Support Vector Machine ou Séparateurs à Vaste Marge) sont un ensemble de techniques d'apprentissage supervisées permettant de traiter des problèmes de discrimination. Elles consistent à représenter les vecteurs à classer dans des espaces de plus grande dimension et à chercher les hyperplans permettant une séparation optimale des classes [GCH99].

[18]Les RVM (Relevance Vector Machines) sont également des méthodes de classification permettant d'obtenir des degrés d'appartenance à une classe en utilisant des méthodes basées sur les inférences bayesiennes. Les données sont représentées dans un espace de plus grande dimension puis on construit un modèle parcimonieux de leur distribution dans ce nouvel espace [Tip00].

d'être capable de la reconstruire à l'aide de la vidéo. Cette approche peut être envisagée dans le traitement de caméra de vidéo-surveillances ou le fond est visible la majorité du temps. Elle est difficilement applicable dans une vidéo en LS ou le signeur occulte la plupart du temps la partie centrale du fond. Lorsque le fond est variable, il est possible de modéliser la distribution de la couleur de chaque pixel de fond sous forme d'une gaussienne [WADP96] ou d'une multi-gaussienne [SG99]. Certaines méthodes utilisent la différence inter-image pour distinguer le fond des objets qui l'occultent. Un pixel qui change peu de couleur pendant un certain nombre d'images est alors considéré comme appartenant au fond [AA01]. Notons que l'utilisation de la différence inter-image est aussi utilisée dans de nombreux travaux comme [HLM04] pour améliorer une segmentation opérée sur un critère de couleur. Une méthode basée exclusivement sur la différence inter-image serait inapplicable pour des vidéos en LS telles que les nôtres car de nombreux pixels situés au milieu des images sur les vêtements du signeur changent relativement peu de couleur au cours de la vidéo et risqueraient donc d'être classés comme fond.
Dans le même esprit, certaines méthodes d'extraction du fond sont basées sur le flot optique entre plusieurs images consécutives. Les zones statiques de l'image sont alors classées comme appartenant au fond. Outre le fait que ces méthodes sont assez coûteuses en temps, les mouvements rapides des LS rendent difficile leur application.

Il peut être intéressant pour des traitements ultérieurs d'appliquer à l'image des **filtres de détection de contours, d'arêtes et de sommets**. Ceci peut en particulier être réalisé en utilisant les détecteurs mis au points par [Can86] ou [SC92]. Les arêtes sont des informations intéressantes car elles permettent de localiser les avant-bras même lorsque ceux-cis sont situés devant le buste du signeur (et donc non-détectables grâce à la silhouette) [DC01].

Le calcul **d'images intégrales**[19] d'images de peau ou d'images de silhouette peut également s'avérer pertinent si les méthodes de traitement ultérieures de la vidéo font appel à un calcul fréquent d'intégrales. Les images intégrales sont en particulier utilisées par [HYDD05] pour effectuer des suivis de personnes et par [LW07] pour effectuer un suivi de la tête.

4.2.7 Modélisation de l'apparence du corps humain

Une fois les prétraitements de la vidéos effectués, il est possible d'effectuer une détection ou un suivi de plus haut niveau du système à suivre. Dans le cadre de la reconstruction de posture, on peut distinguer les approches qui modélisent les différents membres séparéments (principalement les approches *basées modèles*) et celles qui caractérisent l'apparence d'un groupement de membres en comparant l'image obtenue à une base d'apprentissage (approches *basées images*). Comme l'a montré [Fon08], la sélection d'indices visuels pertinents pour effectuer la reconstruction de la posture joue un rôle souvent aussi im-

[19] Le lecteur peut se reporter à l'annexes II pour une description du fonctionnement et de l'intérêt des images intégrales.

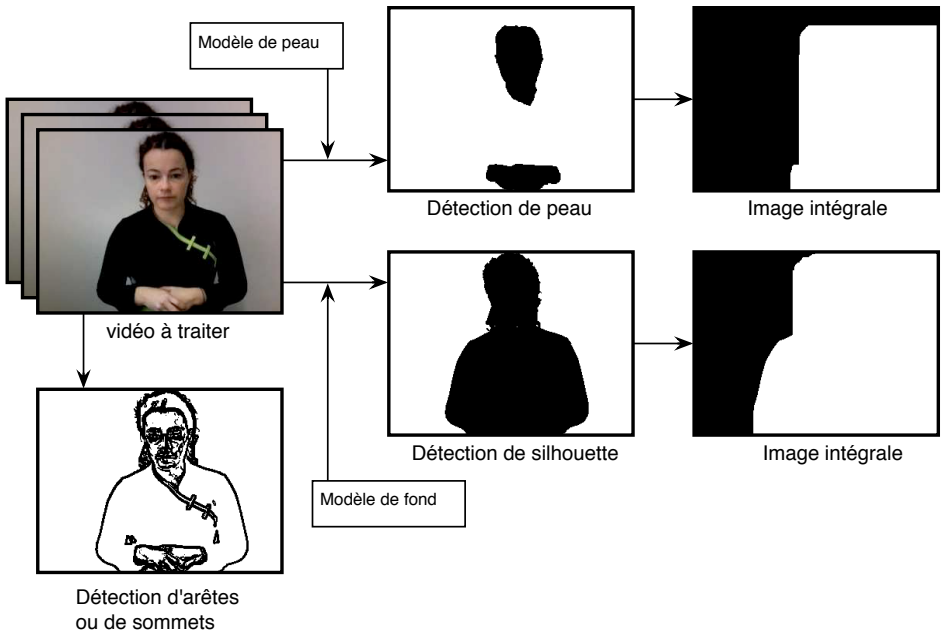


FIGURE 4.7 – Un exemple de prétraitement de vidéo

portant que les méthodes d'optimisations utilisées ultérieurement dans la reconstruction de la posture 3D. Pour cette raison, nous décrirons dans cette partie les différents modèles d'apparence mentionnés dans la littérature en montrant à chaque fois s'ils semblent pertinents pour l'analyse de vidéos en LS. Nous mettons de côté tous les indices de détection ou de suivi basés sur des marqueurs colorés sur les mains [HGO96] ou sur le corps du signeur [YC00] car nous ne faisons pas usage de marqueurs dans nos vidéos.

4.2.7.1 Membres séparés

Une première approche de modélisation est de décrire l'apparence des différents membres séparément. Nous aborderons différents modèles utilisés pour spécifier l'apparence des mains, de la tête, des coudes, des bras et avant bras et du buste. Il est en effet fréquent dans les différentes méthodes d'opérer une telle segmentation du signeur. Nous ne rentrerons pas dans le détail de la modélisation de l'apparence des configurations manuelles et des expressions du visage. Pour chaque membre que nous venons de mentionner, nous listons les différents modèles d'apparence dans un ordre de complexité croissante. Nous proposons ensuite une illustration des différents modèles.

Pour la **détection et le suivi des mains**, le premier indice utilisé est la couleur de la peau, que certaines équipes fusionnent avec la détection de mouvement inter-image pour être plus robuste aux pixels de fond qui auraient une teinte proche de la chair. La recherche de la main peut donc être effectuée sur une image prétraitée telle que l'image de détection de peau de la figure 4.7. Il est toutefois possible d'utiliser également la forme de la silhouette de la main. Suivant les méthodes de suivi, la modélisation de la forme de la main est plus ou moins élaborée.

Etant donné que les mains n'ont pas de forme bien caractérisée, mais que leur morphologie leur impose un encombrement maximal, [BWK+04] propose de les modéliser par des carrés orientés horizontalement et verticalement dans l'image. Cette modélisation est certes plus inexacte que si une forme ovale avait été choisie (on choisit arbitrairement l'orientation des carrés), mais elle présente l'intérêt de mettre en oeuvre des techniques de calculs basées sur les images intégrales très peu coûteuses en temps de calcul.
Pour mieux tenir compte de la forme globale de la main, d'autres approches proposent de les modéliser par des ellipses à l'instar de [SWP98]. Les paramètres des ellipses sont cependant extrêmement dépendants de la précision de la détection des pixels de la main. De plus, l'orientation de l'ellipse dépend à la fois de la configuration manuelle et de l'orientation de la main et est donc très difficile à interpréter pour obtenir l'orientation de la main (même approximative).
Notons que des primitives géométriques tridimensionnelles sont également utilisées pour modéliser l'apparence des mains. Il est possible d'employer par exemple des ellipsoïdes [DC01] ou des cylindres [DBR00]. [Gia08] propose de modéliser la main comme un nuage de particules de couleur peau. Cette modélisation a l'avantage de tenir compte des grandes variations de forme de la main et de ne pas nécessiter de segmentation de la main préalablement au suivi.
[WADP96] modélise les mains comme des blobs (ensemble de pixels connexes) de couleur peau. Là aussi, la méthode s'adapte aux grandes variations de configuration de la main mais nécessite un temps plus important de segmentation du blob dans l'image.

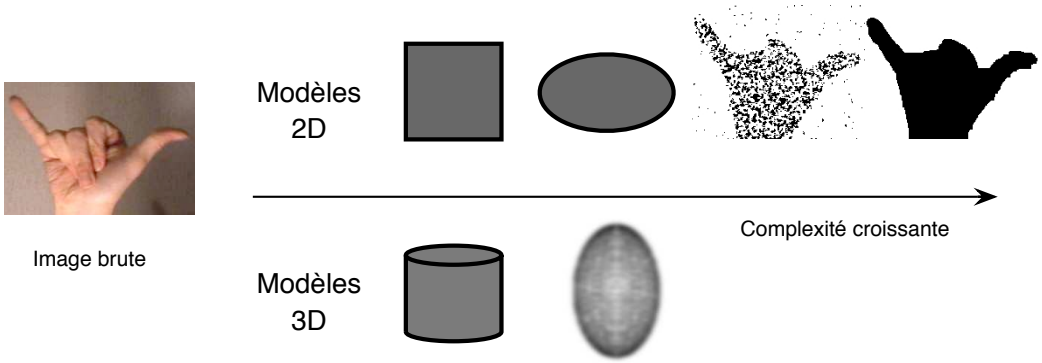


FIGURE 4.8 – Différents modèles d'apparence de la main

Le suivi de la tête utilise également les détections de la peau et de la silhouette du signeur. Par opposition aux mains, la forme de la tête du signeur en projection 2D varie relativement peu dans les vidéos car le signeur a tendance à regarder en direction de la caméra[20]. Il est à souligner que contrairement à la main, la tête est souvent discernable dans les contours de la silhouette. Nous aborderons rapidement les modèles similaires à ceux utilisés pour le suivi de la main pour nous concentrer sur les modèles spécifiques à la tête. Il est naturellement possible d'envisager des modèles grossiers comme la modélisation sous forme de rectangle [JBY96] qui permet d'utiliser les images intégrales [BWK+04]. On notera que dans le cas de la tête,

[20]Sauf en cours de transfert où on note de des rotations de relativement faible amplitude (une trentaine de degrés environ).

les dimensions de ces rectangles varient peu au cours de la vidéo, même si le signeur tourne la tête.
Du fait de la faible variation de l'apparence de la tête, il est possible d'utiliser comme modèle l'image moyenne de sa silhouette comme dans [RMR04]. La modélisation de la tête sous forme d'ellipse prend cette fois tout son sens car la forme de la tête est globalement une ellipsoïde.
On retrouve les modélisations sous forme de primitives 3D de la tête comme les ellipsoïdes [DF99], et les cylindres [DD92].
La tête peut également être modélisée sous la forme d'un blob [HLM04] ou d'un nuage de points [Gia08].
Contrairement aux mains dont la texture est relativement difficile à modéliser (car trop dépendante de la configuration, de l'illumination et de la texture des mains du signeur), la tête du signeur a globalement une apparence peu affectée par les expressions du visage. Ceci rend possible l'élaboration d'autres techniques de détection et de suivi de la tête basées sur la texture. Le filtre de Viola et Jones [VJ01] est une méthode rapide et efficace pour localiser les têtes dans une image. Elle est basée sur des filtres relativement rudimentaires dont le calcul est basé sur des images intégrales. Les résultats des filtres élémentaires sont fusionnés par une méthode de boosting [SFBL98]. Les performances de cette méthode sont toutefois nettement moins satisfaisantes lorsque la tête est occultée. La méthode ne tolère pas non plus les rotations dans le plan.

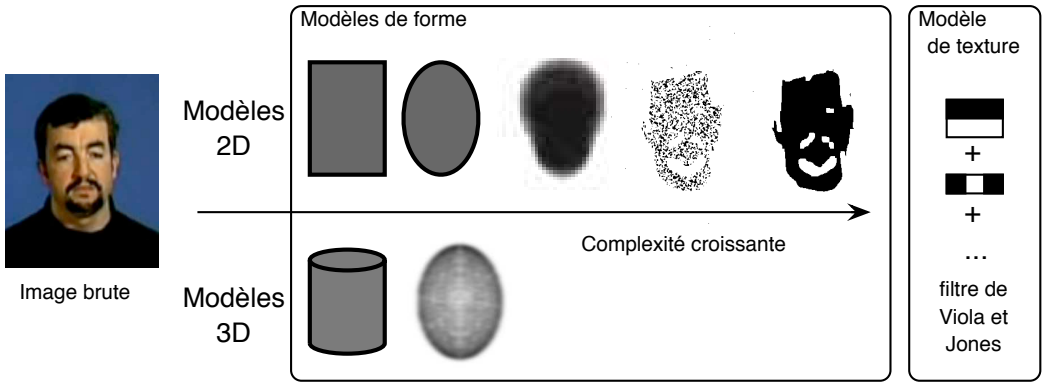


FIGURE 4.9 – Différents modèles d'apparence de la tête

La **détection des bras et des avants bras** du signeur peut se faire soit en se basant sur la couleur de peau (il est alors nécessaire que le signeur ait des manches courtes), soit par l'intermédiaire de la silhouette. Il est important de souligner que dans le cadre des vidéos en langue des signes, l'avant bras est le plus souvent difficile à discerner dans la silhouette du signeur en raison des vêtements unis. Les différents modèles d'apparence utilisés dans le cadre du suivi appartiennent aux catégories qui suivent.
[RMR04] utilise une image moyenne de bras mais émet l'hypothèse que les bras de la personne suivie sont toujours le long du corps. Evidemment, une telle hypothèse n'est pas réaliste dans le cadre des LS où les orientations des bras et des avants bras varient constamment.
Ceci conduit [JBY96] à modéliser les bras et les avants bras du signeur par des rectangles de largeur connue et d'orientation variable. Les modèles 3D utilisent en général des primitives telles que les cônes tronqués [WN97] [DBGP96], les cylindres [DF99] et les ellipsoïdes tronquées [DC01].

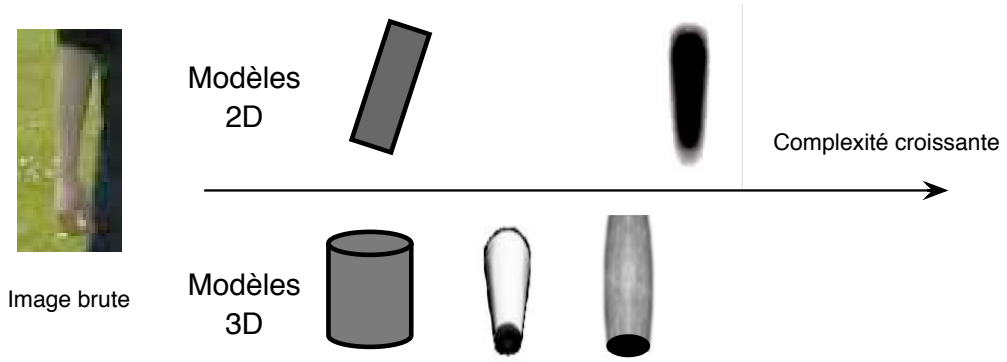


FIGURE 4.10 – Différents modèles d'apparence de la tête

Le buste du signeur peut être suivi grâce à sa silhouette. Malgré la présence de transferts durant lesquels le signeur peut effectuer des balancements et des rotations hors plan de faible amplitude[21], l'apparence du buste varie peu dans la vidéo.
[RMR04] utilise des apparences moyennes de la silhouette pour effectuer le suivi de buste.
[BWK+04] détecte le buste par une méthode basée sur les images intégrales en modélisant le buste comme un rectangle.
Différentes primitives 3D sont utilisées pour modéliser la forme du buste. Parmi elles, on peut citer les cylindres [DF99] et les ellipsoïdes tronquées [DC01].

Peu d'études se concentrent sur le **suivi des épaules et des coudes**. La position de ces articulations est alors déduite à partir de celle des membres adjacents.
Etant donné la grande déformabilité de ces zones d'articulations, des méthodes comme les contours actifs sont bien adaptées à effectuer un suivi des épaules. [Gia08] reporte l'utilisation de contours actifs pour suivre les épaules d'une personne effectuant des allées et venues dans une pièce.
Le suivi des coudes est relativement difficile dans le cas général car l'apparence des coudes est extrêmement variable. Cependant, nous montrerons au §6 que leur apparence varie peu dans une séquence filmée de LS.

Plusieurs méthodes de fusion de détecteurs locaux existent pour mettre au point des détecteurs plus sophistiqués. On citera les méthodes de boosting, de SVM ou de fusions bayésiennes qui permettent de réaliser des fusions linéaires ou non linéaires de détecteurs locaux. [GCG+96] donne un bon exemple de fusion de plusieurs indices (texture, couleur, mouvement) pour détecter la tête du signeur.

Dans la plupart des modèles, la synthèse de la silhouette du corps humain est obtenue par une juxtapo-

[21] Le fait que le signeur soit filmé engendre une diminution importante de l'amplitude de ces rotations hors plan.

sition de la silhouette de chacun des membres. Des modèles de corps humains plus élaborés sont cependant également utilisés dans le cadre de la reconstruction automatique de la posture. Ainsi, [HP91] utilise un modèle déformable de bras et [PF01] emploie un modèle d'humanoïde en modélisant sa peau.

4.2.7.2 Indices plus globaux

Nous avons mentionné dans la partie qui précède, des modèles d'apparence des différents membres, pris séparément les uns des autres. Certaines méthodes de reconstruction de la posture sont basées sur des modèles d'apparence de corps dans son intégralité. La plupart d'entre elles utilisent l'information de silhouette du signeur. [RS00] utilise les moments de Hu [Hu62] pour caractériser la silhouette du signeur. On trouve dans [CADlT04] des transformations robustes permettant une mise en correspondance plus aisée des silhouettes. [HHD98] utilise des transformées de distances entre les différents points de la silhouette du signeur pour identifier les différents membres du signeur. Les auteurs mentionnent cependant dans leur article une hypothèse forte selon laquelle les mains sont la plupart des temps situées en périphérie de la silhouette humaine. Cette hypothèse est la plupart du temps non vérifiée en langue des signes française et nous le montrerons au §6. Suivant une approche apparentée, [AT04] utilise des *contextes de formes* [BM00a] pour faciliter la mise en correspondance de silhouettes humaines.
L'utilisation de contours actifs [JB03] est fréquente pour la caractérisation de silhouette. Ainsi, [BH94] les utilise pour suivre la silhouette de piétons. L'auteur extrait ensuite les coordonnées de points de contrôle de la spline qui approxime au mieux la silhouette. Dans le même esprit, [Gia08] utilise des coutours actifs modélisés par des courbes de Bézier pour suivre la tête et les épaules dans le cadre d'Interaction Homme-Machine.

4.2.7.3 Gestion des occultations et des sorties de cadre

Il est possible de remarquer, en analysant les différents détecteurs locaux couramment utilisés, que des indices similaires sont utilisés pour détecter et suivre différents membres. Ainsi par exemple, la tête et les mains sont fréquemment localisés grâce à des filtres utilisant un modèle de couleur de peau. Ceci pose à la fois des problèmes de gestion d'occultations (que se passe-t-il par exemple lorsqu'une main passe devant la tête ?) et de désambigüisation (comment peut-on distinguer la main gauche de la main droite ?).

Pour résoudre ou contourner le problème des occultations, plusieurs approches sont proposées dans la littérature :

- [ILI98] utilise le fait que la tête a un mouvement faible pour identifier correctement la main et la tête du signeur. Cette hypothèse est partiellement vérifiée dans le cadre de la LSF, mais donne tout de même des résultats satisfaisants.

- Certains auteurs comme [Gia08] utilisent des modélisations qui tolèrent la superposition de plusieurs membres dans l'image. Dans le même temps, un mécanisme d'interaction entre les détecteurs permet

de faire en sorte que plusieurs détecteurs ne suivent pas le même membre. C'est aussi l'approche envisagée par [MI00] qui utilise un modèle dans lequel chaque pixel peut appartenir à une ou l'autre des parties du système d'apparence similaire.

- D'autres travaux comme [NB07] mettent en oeuvre des forces de répulsion entre les mains de la personne dont on souhaite reconstruire la posture, de manière à ce que les mains soient le plus possible séparées. Cette approche, bien qu'aussi employée pour suivre des mouvements de LSF, n'est pas souhaitable car elle nie l'importance des contacts qui jouent un rôle important dans cette langue.

- Une autre approche est de modéliser les mains et la tête comme trois blobs séparés et de les fusionner lorsqu'il y a occultation. C'est l'approche *Merge and Split* proposée par [HCZ08]. Le problème de cette approche est qu'on ne peut alors pas localiser précisément chaque membre lors de l'occultation.

Un autre problème lié à celui que nous venons d'évoquer est celui de sortie de cadre de certains membres. Ce problème est relativement peu abordé dans les articles traitant de la reconstruction de la posture, car les conditions de capture du mouvement permettent de voir l'ensemble des membres supérieurs du signeur. Dans le cas des vidéos en LSF que nous traitons, il est extrêmement fréquent que les mains du signeur sortent en bas du cadre de la vidéo. Les détecteurs locaux ont alors tendance à affecter la main disparue à la partie de l'image qui se rapproche le plus visuellement du membre à suivre.

Une fois que les deux mains et la tête sont correctement localisées dans l'image, se pose un troisième problème de désambigüisation. En effet, si la tête et les mains ont des morphologies différentes qui permettent de les identifier. Les deux mains sont extrêmement ressemblantes au niveau de la forme, de la couleur et de la texture. Plusieurs solutions sont proposées dans la littérature :

- La première que nous écartons d'emblée consiste à contraindre les mouvements du signeur de manière à ce que la main droite soit toujours à droite de la main gauche (du point de vue du signeur). Cette solution est irréaliste pour les LS dans lesquelles les mains peuvent être "inversées" plus de 10% du temps[22].

- Lorsque le signeur porte un vêtement à manche courte, le problème d'identification des mains est bien plus aisé car il découle directement de l'identification du bras droit et du bras gauche qui sont alors détectables sur l'image de peau. Lorsque le signeur porte des manches longues en revanche, il est très délicat d'utiliser les caractéristiques des mains. Les techniques de désambiguisation que nous avons mises au point pour l'identification des mains d'un pianiste [HLA09] basées sur les directions principales des ellipses utilisées pour modéliser les mains donnent des résultats inexploitables pour les vidéos en LS. La raison de cet échec est que les changements de configurations manuelles rendent le modèle d'ellipse inadapté à notre cas d'application.

[22]Mesure effectuée sur une traduction de brève d'actualité en LSF de 30s fournie par la société Websourd.

- Il est donc indispensable de prendre en compte d'autres critères pour pouvoir opérer cette identification. Des approches comme les filtres relationnels [JJC02] permettant de modéliser les liaisons entre les différents membres pourraient apporter une solution au problème.

- Finalement, l'utilisation d'une base d'apprentissage sur l'apparence peut être utilisée. Ainsi, [HCZ08] utilise les moments de Hu des blobs de peau pour inférer des hypothèses sur les positions de la main et de la tête dans, ou hors de l'image. L'intérêt de cette approche est de résoudre en même temps les problème d'occultation, de sortie de cadre et de désambiguisation.

4.2.7.4 Conclusion

Nous avons décrit différents indices visuels qui peuvent être utilisés pour effectuer une identification puis un suivi des différents membres du corps du signeur. La reconstruction de la posture dans le cadre de la LSF offre un défi majeur en raison du nombre important d'occultations entre les différents membres. Il est donc évident que dans un contexte de monovision, l'utilisation d'un seul des indice (carte de peau ou silhouette) ne peut pas conduire à une reconstruction satisfaisante de la posture. C'était la conclusion des auteurs de [WS03] qui déploraient que peu d'équipes produisent des solutions vraiment performantes et abouties. Depuis, les méthodes de suivi ont évolué prenant en compte de plus en plus d'indices et les fusionnant à la manière de [Fon08] et [NB07]. Nous nous inscrivons également dans cette démarche.

4.2.8 Modélisation de la dynamique du corps humain

4.2.8.1 Modèles géométriques du corps humain

Les approches d'associations explicites entre les résultats obtenus à l'aide du suivi des membres et les postures 2D ou 3D correspondantes font en général appel à un modèle cinématique du corps humain. Il s'agit donc d'approches *basées modèle*. En général, les modèles proposés représentent le corps humain comme un système articulé de solides indéformables.
Chaque méthode doit choisir un modèle cinématique assez riche pour retranscrire fidèlement le mouvement et y ajouter éventuellement des contraintes reflétant connaissances *a priori* sur le système (butées articulaires, liaisons cinématique ou rigidité des articulations). Il est alors fréquent que le degré de liberté du système soit inférieur à la somme des degrés de liberté de chacune de ses articulations.
Plusieurs aspects doivent être envisagés pour choisir la richesse du modèle cinématique de corps humain :

- Le nombre de degrés de liberté du système doit être assez élevé pour fournir une estimation de la posture utilisable dans le contexte d'application envisagé. Dans notre cas, le modèle devra au moins être compatible avec un positionnement des mains en tout point de l'espace de signation pour permettre de retranscrire fidèlement les mouvements impliqués dans les signes.

- Le degré de liberté utilisé pour la reconstruction du système est de toute façon limité par la quantité d'information disponible. Ainsi, la méthode décrite par [BWK$^+$04] permet de déterminer la position

la plus probable des coudes à partir de celle des mains, du buste et de la tête mais ne pourra en aucun cas permettre de reconstruire n'importe quelle posture 3D du bras, bien que les degrés de mobilité de chaque articulation semblent être respectés.

- Un nombre important de degrés de liberté du système implique un temps de calcul important pour estimer la posture qui correspond le mieux aux observations de l'image. Cette augmentation du temps de calcul pour le parcours de l'espace des solutions est appelé *fléau de la dimension* (curse of dimensionnality) selon l'appellation de Richard Bellman [Bel03].

Une fois le choix effectué sur la chaîne cinématique et sur le nombre de degrés de liberté que nous souhaitons lui laisser, il reste à savoir comment ces modèles peuvent être exploités. Il est possible à la fois de tirer parti des inégalités qui découlent des modèles comme l'éloignement maximal de plusieurs membres en projection 2D et d'utiliser des équations de cinématique inverse.

Dans de nombreux cas, les simples modèles cinématiques conduisent à des équations n'admettant pas une unique solution. Il suffit pour s'en convaincre de prendre l'exemple d'un bras dont on connaîtrait uniquement les coordonnées du coude (x_c, y_c) et de l'épaule (x_e, y_e) dans l'image de la vidéo (cf. image 4.11). Si la caméra est correctement calibrée, il est possible de retrouver les coordonnées (X_c, Y_c),(X_e, Y_e) dans l'espace réel à partir de leurs coordonnées (x_c, y_c) et (x_e, y_e) dans l'image[23]. Si on connaît en plus la longueur réelle L du bras du signeur, il est possible d'estimer ΔZ, la différence de profondeur entre le coude et l'épaule du signeur.

$$\Delta Z = \pm\sqrt{L^2 - (X_c - X_e)^2 - (Y_c - Y_e)^2}$$

Ce problème est mentionné dans deux nombreuses publications comme dans [Gia08]. Il est alors nécessaire d'utiliser d'autres hypothèses sur les postures pour lever les ambiguïtés. Ces contraintes peuvent être par exemple des contraintes sur les articulations comme des angles limites ou des rigidités d'articulation comme dans [JBY96].

4.2.8.2 Modèles implicites d'association observation-posture

Par opposition aux modèles géométriques précédents basés sur une modélisation explicite du corps humain sous la forme d'une chaîne cinématique, d'autres techniques que nous allons exposer permettent de faire une association observation-posture grâce à un modèle implicite basé sur un apprentissage.
La problématique du nombre de degrés de liberté du système reconstruit demeure. Un système avec trop peu de degrés de liberté ne donne pas une reconstruction fidèle du mouvement. Inversement, un système avec trop de degrés de liberté nécessitera une base d'apprentissage conséquente permettant une couverture

[23]Nous prenons ici pour simplifier l'hypothèse que la caméra se trouve suffisamment loin du signeur pour qu'on puisse considérer qu'il y a un simple changement d'échelle entre les coordonnées (X,Y) de l'espace réel et les coordonnées (x,y) de l'image.

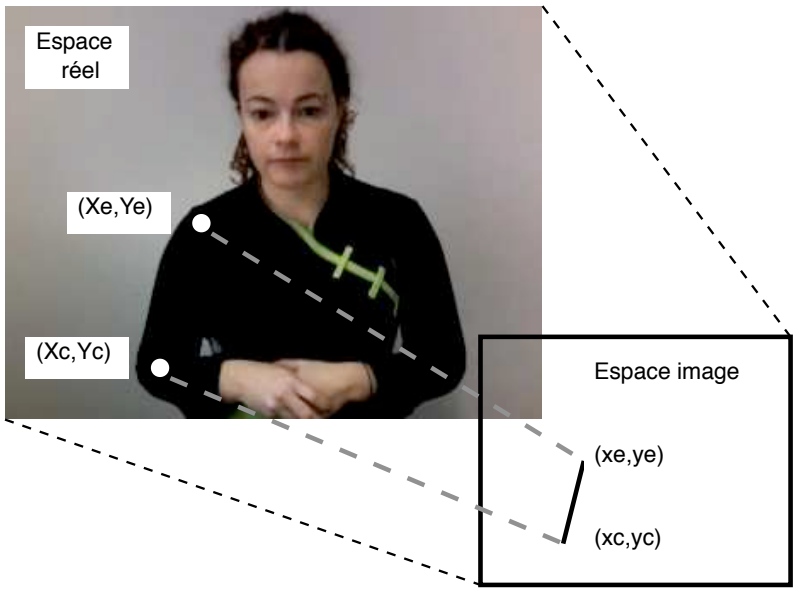


FIGURE 4.11 – Détermination de la différence de profondeur entre l'épaule et le coude par cinématique inverse

satisfaisante de l'espace de postures représentées. Le problème d'association d'observation et de posture peut être formulé de la manière suivante : On dispose d'un vecteur d'observations dans l'espace d'observations et on cherche à l'associer à un vecteur d'angles (ou de coordonnées spatiales) dans un espace représentant les postures. On cherche donc une application permettant de faire le lien entre ces deux espaces.
Le vecteur d'observation peut contenir des données extrêmement variées. Nous n'exposons ici que quelques exemples représentatifs :

- Les formes géométriques peuvent y être représentées par leurs paramètres (coordonnées 2D, dimensions, orientation, moment de Hu [Hu62] ...),
- Les contours peuvent être représentés à l'aide des coordonnées de points de contrôles comme dans [Gia08] et [BH94],
- Les silhouettes peuvent aussi être modélisées par des *contextes de forme* comme dans [BM00a].

L'apprentissage d'exemples d'associations entre vecteurs d'observation et vecteurs de posture permet alors de construire implicitement l'application qui permettra de passer du vecteur d'observation au vecteur de posture. Nous détaillons dans les lignes qui suivent différentes approches pour construire cette application :

- Le modèle d'association peut être linéaire. C'est en particulier le cas lors de l'utilisation d'ACP comme dans [BH94] ou [Gia08]. Le degré de liberté du système reconstruit est alors limité par le nombre de composantes principales utilisées. Cette méthode est assez satisfaisante sur un plan

théorique mais les hypothèses qu'elle sous-tend sont loin d'être vérifiées. L'espace des postures ainsi que l'espace des observations ne sont pas forcément bien représentés par une combinaison linéaire de postures (respectivement d'observations) de base. L'approximation linéaire de la transformation entre l'espace d'observations et l'espace de postures n'est valable que sur des domaines restreints. Il n'est qu'a considérer les équations impliquées dans la cinématique inverse telles que celles du §4.2.8 faisant intervenir des racines carrées pour s'en convaincre.

- Il est naturellement possible d'utiliser des variantes non linéaires des ACP pour effectuer l'association entre observations et postures. L'article [DDR+05] montre un exemple de généralisation des ACP en utilisant des noyaux polynomiaux et montre une application pour retrouver les points caractéristiques du crâne à partir d'une image de crâne, qui pourrait être généralisable pour notre problème de reconstruction de la posture.

- Par opposition à ces modèles d'associations valables sur tout l'espace des observations, d'autres applications peuvent être définies localement et différer en fonction de la portion de l'espace des observations. C'est le principe des méthodes d'interpolation de la posture à partir des plus proches voisins dans l'espace des observations. La recherche des plus proches voisins peut être rendue plus rapide en utilisant des tables de hachage [PVD03] ou des arbres de recherche [SBS02] de manière à limiter le temps de reconstruction de la posture pour chaque image.

4.2.8.3 Modèles de mouvement

Nous cherchons à reconstruire la pose d'un signeur à partir d'une vidéo. Il est donc possible de tirer également parti de modèles de mouvement pour rendre plus précises nos estimations de posture 3D.
[AT04] propose d'utiliser un modèle linéaire de mouvement dans son filtre de Kalman. La position prédite d'un membre peut alors être estimée comme une combinaison linéaire des deux emplacements précédents de ce membre. L'auteur constate qu'un simple prédicteur basé sur une hypothèse de vitesse constante conduit à des oscillation dans la reconstruction de la posture. Il préconise donc l'utilisation d'un prédicteur utilisant un modèle de mouvement amorti. Bien que l'hypothèse d'un mouvement à vitesse constante puisse être étayée d'un point de vue statistique pour des mouvements de LS, il ne nous semble pas pertinent de l'employer dans ce cadre. En effet, ces prédicteurs peuvent conduire à des décrochages lors de points de rebroussement de la trajectoire. Pour ces raisons, [SG00] suggère d'utiliser un simple modèle de marche aléatoire pour la prédiction de trajectoire. Ce modèle de mouvement permet notamment de tenir compte de la continuité du mouvement. La raison du succès mitigé de modèles de mouvements à vitesse constante s'explique par le fait qu'il existe plusieurs *régimes* de dynamique durant la productions de signes (mouvement accéléré, mouvement amorti, mouvement à vitesse constante, arrêt brusque ...). Pour tenir compte de cette constatation, les auteurs de [PRC00] proposent d'utiliser successivement plusieurs modèles de dynamiques en fonction des observations obtenues lors du traitement de la vidéo.

4.2.9 De l'observation à la reconstruction : obtention de la posture du signeur

Les différentes méthodes de reconstruction de la posture doivent maintenant utiliser au mieux les différents modèles que nous venons d'évoquer pour reconstruire la posture du signeur :

- Modèle d'apparence,
- Modèle cinématique du corps humain,
- Modèle d'association implicite observation-posture,
- Modèle de mouvement.

Ce problème revient souvent à minimiser une fonction de coût.

4.2.9.1 Quelle fonction de coût optimiser ?

Le but des fonctions de coût optimisées dans le cadre de la reconstruction de la posture est de maximiser la ressemblance entre les images de la vidéo et les images qui pourraient être générées d'après les postures estimées et les modèles d'apparence (explicites ou implicites). Elles résultent en général de la combinaison de plusieurs mesures de ressemblances ou de distances. La liste que nous présentons n'est en aucun cas exhaustive mais nous reprenons les principales mesures de similarité utilisées dans le cadre de la reconstruction 3D de la posture :

- Si les contours sont utilisés, un des critères peut consister à minimiser la distance moyenne entre les contours de l'image de la vidéo et les contours qui seraient générés par la posture du signeur d'après le modèle d'apparence. Ce critère est notamment utilisé par [DC01] et [DF99].
- Si les régions sont utilisées, il est possible de minimiser la distance de Hamming entre la silhouette du signeur dans la vidéo et la silhouette qui serait générée par la posture d'après notre modèle d'apparence [SVD03]. C'est un critère que nous utilisons dans nos propres travaux.
- Il est naturellement possible d'utiliser d'autres mesures de distance ou de similarité comme les distances entre *contextes de formes* comme dans [MM02].
- Il est aussi possible d'évaluer le confort de la posture [YTK+04] ou la conformité de la posture obtenue aux différentes contraintes cinématiques comme l'éloignement entre les membres [RMR04]. Cette contrainte peut être interprétée comme une minimisation de l'énergie potentielle de la posture.
- Naturellement, il peut également être intéressant de prendre en compte la dynamique du système dans la fonction de coût. Nous noterons toutefois que la dynamique du système est plus souvent prise en compte indirectement dans les algorithmes d'optimisation. En effet, contrairement aux mesures précédentes, la prise en compte de la dynamique du système fait également intervenir l'état du système à plusieurs instants.

Un critère d'optimisation pris isolément n'est en général pas suffisant. Des études comme [DD92] montrent d'ailleurs que l'utilisation d'un seul type d'indice peut conduire à un manque de robustesse et à des "décrochages" dans les algorithmes de suivi. Il est donc nécessaire de fusionner les différentes mesures de similarité. Plusieurs méthodes de fusion sont mentionnées dans la littérature. Parmi elles, on peut citer la fusion bayésienne naïve ou pondérée et les moyennes pondérées. On trouve rarement des justifications des méthodes de fusion utilisées par les différents algorithmes de reconstruction de la posture. Les auteurs de [YTK+04] montrent que la méthode de fusion utilisée a une influence moindre par rapport aux indices pris en comptes dans la fonction coût.

4.2.9.2 Méthodes d'optimisation

Il est alors nécessaire de trouver quelles postures correspondent le mieux aux images de la vidéo. Nous distinguerons différentes approches pour parvenir à cette optimisation.

Les premières familles d'approches estiment pour chaque image de la vidéo la posture qui minimise la fonction coût. Ceci peut être obtenu à l'aide d'un schéma de Levenberg-Marquardt comme dans [PF01]. La posture de départ est alors initialisée avec la posture précédente, puis le schéma converge vers la posture correspondant à l'image courante au fil des itérations.
De nombreuses autres méthodes d'optimisation itératives utilisent aussi la structure du modèle cinématique du signeur. Ainsi, les travaux [MT93], [DC01] et [BM98] proposent de remplacer les écarts entre la silhouette de l'image courante et la projection de la posture estimée par des forces s'exerçant sur la chaîne cinématique du modèle. A chaque itération de ces algorithme, chaque point de la silhouette projetée est associé à un point de la silhouette de l'image réelle. Les forces à appliquer sur la chaîne cinématique sont calculées en fonction des distances entre les silhouettes. Puis la chaîne cinématique se déforme sous l'action de ces forces. Ce schéma d'optimisation nommé *Iterative Closest Point* est décrit dans [Zha94].
Ces premiers types de méthodes posent plusieurs problèmes important. Ils n'utilisent l'information de dynamique du système que par l'initialisation de la posture à déterminer par la posture précédente. Les schémas d'optimisation pas à pas sont également susceptibles de converger vers des minima locaux de la fonction coût. Enfin, il est important de souligner qu'il peut exister plusieurs solutions pour la posture 3D du signeur qui sont aussi vraisemblables pour une même image 2D, ne serait-ce que parce que les équations de la dynamique inverse admettent en générale plusieurs solutions.

Une deuxième famille de méthodes maintient en permanence plusieurs hypothèses sur la posture du signeur à reconstruire.
En réponse au problème de solutions multiples de cinématique inverse, [ST03] propose d'utiliser la structure du problème en représentant les différentes postures correspondant à l'image courante par un arbre. L'utilisation d'informations comme les butées articulaires peut permettre de supprimer certaines branches de cet *arbre des postures* au cours du suivi. Ce schéma d'optimisation basé sur la programmation dynamique est intéressant car il permet d'utiliser à la fois les informations des images précédant et suivant

l'image courante pour déterminer la posture courante.
[CR99] propose de d'étendre l'utilisation des filtres de Kalman pour propager simultanément plusieurs hypothèses. Il remplace donc les distributions gaussiennes traditionnellement utilisées dans les filtres de Kalman par des distributions multi-gaussiennes.
Pour maintenir un nombre important d'hypothèses à chaque instant, il est possible d'utiliser des filtres particulaires[24]. Ainsi [GPS+06] utilise ce type de filtres pour reconstruire la posture d'une personne. L'utilisation de phases de recuit comme dans [DBR00] permet au filtre de converger plus rapidement, de nécessiter un nombre moins important de particules et d'éviter les minima locaux. L'utilisation de phase de recuit, bien qu'elle donne des résultats satisfaisants est discutable sur un plan théorique car elle minimise l'utilisation des informations liées à la dynamique du système. D'autres améliorations des filtres particulaires comme le partitionnement du filtre particulaire et la génération pseudo-aléatoire du nombre de particules permettent d'obtenir une efficacité similaire du filtre, tout en diminuant son nombre de particules. Pour une comparaison des différents filtres particulaires qui peuvent être utilisés dans le cadre du suivi, nous invitons le lecteur à se reporter à [Fon08].

Notons que dans les méthodes basées sur l'utilisation de modèles implicites d'association observation-posture, l'obtention de la posture peut souvent être directe. Ainsi, la recherche de la meilleure posture peut être ramenée à une combinaison linéaire de postures correspondant aux plus proches voisins de l'observation courante, à une combinaison linéaire de postures correspondant à des composantes principales, ou à un calcul direct d'une posture à partir d'un vecteur d'observation, d'après une fonctions apprise grâce à une base d'apprentissage.

4.3 Conclusion

Nous avons pu aborder au fil de cet état de l'art de nombreuses méthodes de reconstruction de la posture, en nous concentrant sur les méthodes monovues. Depuis ces dix dernières années, on note d'importantes évolutions dans les méthodes développées. Il suffit pour s'en convaincre de comparer les travaux actuels à la situation de la recherche présentée dans les états de l'art [Gav99] et [MG01]. Les deux tendances les plus importantes sont :

- L'utilisation de modèles de plus en plus riches sur l'apparence visuelle et la description cinématique du corps humain (modèles cinématiques, rigidité des articulations, confort de posture),
- La mise en place quasi systématique de méthodes qui conservent à chaque instant plusieurs hypothèses sur la posture réelle du signeur (par le biais de distributions multimodales ou particulaires).

Par contre, les solutions originales proposées pour modéliser la dynamique des mouvements sont à la fois peu nombreuses et mal adaptées aux LS [SG00].

[24] Nous invitons le lecteur intéressé par le fonctionnement de ces filtres à se reporter aux annexes II pour obtenir plus de détails techniques.

Reste le problème de l'évaluation des différentes méthodes. Il est extrêmement difficile d'en proposer une comparaison car les mouvements utilisés sont extrêmement divers dans leurs fonctions (mouvements de marche, gestes coverbaux, langues des signes), et leur vitesse d'exécution est variable d'une étude à l'autre.
Les métriques utilisées sont assez différentes, ce qui rend difficile une comparaison quantitative. A titre d'exemple, [AT04] propose une évaluation basée sur l'erreur moyenne sur l'estimation des angles et [Fon08] donne une erreur moyenne sur l'estimation de la position 3D des différents membres.
Même si ces métriques permettent de quantifier les amélioration qu'il est possible d'obtenir en modifiant les différentes méthodes de reconstruction de la posture, elles sont difficilement interprétables d'un point de vue appliquatif.

CHAPITRE 5

MÉTHODES DE TRAITEMENT AUTOMATIQUE DES LANGUES DES SIGNES

Après avoir vu de quelle manière il était possible d'estimer les postures du signeur, se pose la question de l'exploitation de ces données de capture du mouvement pour une interprétation de plus haut niveau. Ce domaine a fait d'extraordinaires progrès au cours des dix dernières années. Il suffit pour s'en convaincre de considérer l'évolution des travaux entre les deux états de l'art présentés dans [PSH97], puis dans [OR05]. Ces progrès résultent à la fois de l'utilisation d'algorithmes plus perfectionnées et de modèles prenant mieux en compte les spécificités des mouvements..

Le problème du passage d'une représentation sous forme de mouvement à une représentation de langage concerne le Traitement Automatique des Langues (TAL). La LSF étant une langue orale, le Traitement Automatique des Langues des Signes (TALS) est plus proche du Traitement Automatique de la Parole (TAP) que de celui des langues écrites. Toutefois, étant donné que les LS n'admettent pas de formes écrites, il est plus difficile d'effectuer la dissociation du traitement de l'oral et de l'écrit.

La science du TALS est beaucoup plus jeune que le TAP. Pour cette raison, de nombreux travaux appliqués aux LS prennent leur ancrage dans des modèles développés pour les langues vocales. Nous commencerons par exposer sommairement les différentes méthodes utilisées pour le traitement automatique de la parole, en focalisant notre analyse sur les motivations et les hypothèses sous jacentes des différents modèles. Nous montrerons ensuite comment ces méthodes ont été adaptées aux spécificités des LS. Pour finir, nous mettrons l'accent sur les problèmes qui subsistent encore malgré l'adaptation de ces modèles, en axant notre analyse sur le problème de la modélisation de la variabilité des LS au niveau lexical.

La majorité des études que nous citerons sont relatives au traitement d'énoncés en LS capturés sous forme vidéo ou par d'autres dispositifs de capture de mouvement. Nous ferons toutefois appel à d'autres domaines connexes dont les méthodes pourraient être appliquées au traitement des LS, comme la reconnaissance automatique de mouvements ou la commande gestuelle.

Nous prenons le parti de présenter, dans les sections qui suivent, à la fois des méthodes traditionnellement dédiées à la détection et d'autres utilisées pour la reconnaissance bien qu'il s'agisse de problèmes distincts. Nous estimons en effet que des méthodes de reconnaissances peuvent être utilisées pour la détection ou la comparaison de signe (cf. §7). Inversement, les méthodes de détection peuvent être utilisées par des algorithmes de reconnaissance.

5.1 Approche traditionnelle du traitement automatique de la parole

Dans cette première partie, nous présentons sommairement les méthodes les plus répandues pour le Traitement Automatique de la Parole. Notre but est de souligner les différentes étapes de ce traitement et d'analyser les motivations des méthodes mises en œuvre. Ceci nous permettra ensuite de voir en quoi ces méthodes peuvent être ou non adaptées pour le Traitement Automatique des Langues des Signes.

5.1.1 Problème de représentation du signal sonore

Le premier problème à résoudre est celui de la représentation du signal. Dès cette étape, il est nécessaire d'intégrer la variabilité des unités à reconnaître. Dans le cas des langues vocales, le signal sonore est issu d'une vibration des cordes vocales, modulée par un système articulatoire. Les analyses font ressortir plusieurs composantes de ce signal :

- La fréquence fondamentale,
- L'énergie du signal,
- Le spectre de fréquence du signal.

Il s'agit de trouver une transformation à même de dissocier ces différents paramètres les uns des autres. Nous citerons à titre d'exemple le Codage Prédictif Linéaire [Bra00], l'utilisation de coefficients cepstraux [HVM96] et le taux de passage à zéro du signal sonore. Ces transformations appliquées à des fenêtres temporelles de 10 à 20 ms permettent de représenter le signal dans un espace continu de plus petite dimension (quelques dizaines de composantes en générale), et de caractériser les différents types de phonèmes composant les langues vocales.
Nous ne rentrons pas dans le détail des traitements mathématiques utilisés pour la caractérisation du signal vocal qui ne serait pas transposable pour les LS, mais nous tenons à souligner le fait que la représentation initiale du signal à traiter est un facteur déterminant pour la réussite de la reconnaissance de la parole.

5.1.2 Dynamic Time Warping

Une des premières méthodes pour identifier des mots appartenant à des corpus de faible taille a été la comparaison des unités à reconnaître avec un ensemble d'exemples préalablement appris.
En général, le débit de prononciation composant les mots à reconnaître peut varier d'une fois sur l'autre. Pour cette raison, il est nécessaire d'utiliser un algorithme d'alignement des séquences temporelles comme l'algorithme de Déformation Temporelle Dynamique présenté en annexes III.1.
L'inconvénient de telles méthodes est qu'elles nécessitent un temps de calcul important et que les modèles utilisés ont tendance à être très volumineux. Il s'agit d'un des facteurs qui a motivé la mise au points de modèles plus synthétiques comme les Modèles de Markov Cachés, pouvant être utilisés pour effectuer des tâches de reconnaissance.

5.1.3 Les Modèles de Markov Cachés

Les Modèles de Markov Cachés (MMC) sont présentés en détail dans [Rab89], nous en rappelons les fondements dans l'annexe III.2. Ces modèles permettent de représenter un phénomène comme une succession d'états générant aléatoirement des observations discrètes. Dans le cadre de Traitement Automatique

de la Parole, ces observations correspondent en général à des catégories de vecteurs. Une étape de quantification vectorielle est nécessaire pour constituer les catégories. Pour la reconnaissance de mots, chaque état peut modéliser un phonème ou une unité acoustique de plus petite taille. Le schéma 5.1[1] montre la structure d'un modèle de Markov associé au mot "trois". Les phonèmes composant les mots à reconnaître

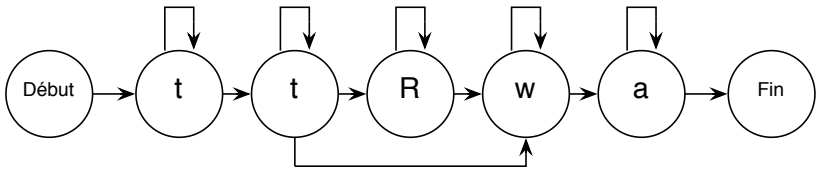


FIGURE 5.1 – Structure simplifiée du Modèle de Markov associé à la prononciation du mot "trois"

peuvent être prononcés de manières différentes suivant les autre phonèmes auxquels ils sont juxtaposés. Les modèles de Markov tiennent compte de ce phénomène de coarticualtion en modélisant des modèles de bigrammes ou de trigrammes. Ainsi, on pourra définir plusieurs états différents correspondant au phonème [y] suivant qu'il se trouve après la consonne [t] ou la consonne [n]. Il peut aussi arriver que des phonèmes soient ajoutés pour effectuer des liaisons entre les mots. Ainsi, la concaténation du mot "un" et du mot "arbre" sera prononcé [œnaRbR]. Ce problème de phonèmes à insérer entre les unités à reconnaître est systématiquement présent en LS, où des transitions sont ajoutées entre deux signes consécutifs.

Il faut savoir que les Modèles de Markov peuvent être utilisés à différents niveaux. S'il est possible de les employer pour modéliser des phonèmes et des mots, leur emploi est également fréquent pour modéliser des langues à partir de statistiques sur les fréquences d'enchaînement des mots qui les composent. Ces modèles statistiques de langues viennent souvent en complément des modèles de lexique dans des systèmes de Traitement Automatique de la Parole.

5.1.4 Les réseaux neuronaux

Il existe également d'autres méthodes de reconnaissance automatique de la parole basées sur des réseaux de neurones construits par analogie avec le fonctionnement humain. Chaque neurone reçoit en entrée un certain nombre de valeurs x_i provenant d'une source extérieure ou d'autres neurones et fournit une valeur de sortie y. Cette valeur est obtenue en appliquant une transformation non linéaire F à une combinaison linéaire des entrées x_i [MP88]. Les neurones peuvent être organisés dans des réseaux à plusieurs couches comme les perceptrons multicouches (cf fig. 5.3). Les perceptrons effectuent une reconnaissance de signaux statiques. Il peuvent ainsi être utilisés pour la classification de certains phonèmes et se présenter comme une étape alternative à la quantification vectorielle comme dans [FSC93].

[1] Cette illustration est inspirée du site http ://alexandre.alapetite.fr/dess-irr/tap/tap02/chiffres.gif

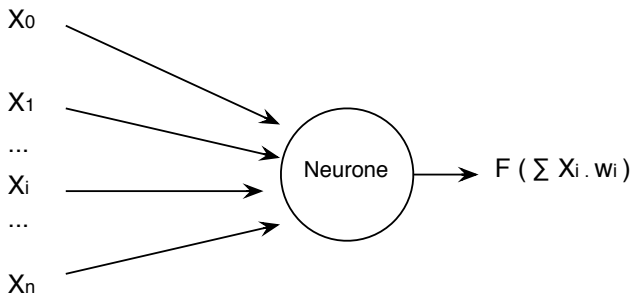


FIGURE 5.2 – Représentation schématique d'un neurone

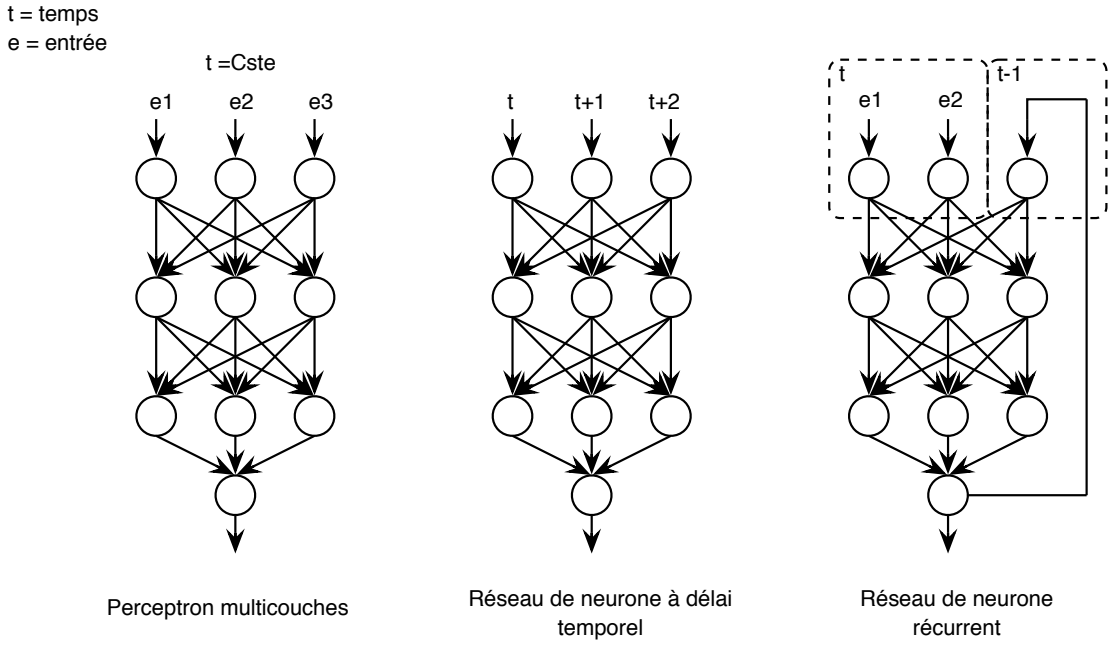


FIGURE 5.3 – Différentes topologies de réseaux neuronaux

Un des inconvénients de l'utilisation des perceptrons pour la reconnaissance des mots parlés est le fait qu'ils traitent mal la dimension temporelle du signal. Plusieurs solutions ont été proposées à ce problème.

- Chaque neurone peut recevoir une caractéristique à un instant précis du signal à reconnaître. On est ainsi dans le cas d'un réseau multicouches à retard (TDNN). Ce type de réseau donne des résultats de reconnaissance satisfaisants pour la détection de mots de faible longueur et de syllabes [WHH$^+$90].

- Il est possible d'utiliser une étape de déformation temporelle dynamique ou d'alignement temporel à l'aide de MMC pour normaliser temporellement les signaux à reconnaître avant de les introduire dans les réseaux neuronaux de types TDNN.

- La modification de la topologie des réseaux neuronaux est aussi une solution pour prendre en compte la temporalité du signal. Dans de tels réseaux, les sorties de certains neurones à l'instant t peuvent être

utilisées comme entrées du réseau à l'instant $t+1$. [Jor90] Ces types de réseaux nommés "réseaux de neurones récurrents" (RNN) peuvent être utilisés pour le traitement de syllabes et de mots de courte durée. Le traitement d'unités plus longues posent des problèmes d'apprentissage des paramètres du réseau de neurone et s'avère long à mettre en oeuvre. Même si des études comme [OR08] portent aujourd'hui encore sur la reconnaissance de la parole exclusivement à l'aide de réseaux neuronaux, il semble que cette méthode soit mal adaptée à la reconnaissance d'un grand nombre d'unités lexicales. Malgré tout, les réseaux neuronaux peuvent être utiles en complément d'autres méthodes comme celles basées sur les MMC et les DTW pour effectuer des tâches telles que la classification de phonèmes [FSC93] ou la suppression de bruits de fond [TW88].

5.2 Difficulté d'application des méthodes traditionnelles de TAL au TALS

Les méthodes de traitement issues du traitement automatique de la parole sont difficilement transposables à la LSF. Nous parlons dans la section suivante des problèmes directement liés aux spécificités des LS, dont certaines sont résumées dans [Bos02] et [DSD+08]. Nous dégagerons les difficultés liées à l'hétérogénéité des données à traiter, à la variabilité des signes, à l'iconicité et la grammaire spatiale.

5.2.1 Des données d'entrées hétérogènes

La première difficulté à surmonter pour parvenir à un Traitement Automatique de Langue des Signes prenant en compte les différents paramètres impliqués dans la production de signe, est de traiter des donnée d'entrées très hétérogènes :

- La posture du signeur peut être représentée comme un vecteur d'angles correspondant aux orientations relatives des différents membres.
- Les configurations manuelles peuvent elles aussi être représentées comme des vecteurs d'angles. Une vingtaine d'angles est en général nécessaire pour les décrire précisément.
- Les expressions du visages sont couramment représentées comme des déformations de texture par les traiteurs d'images[2]. Les modèles peuvent alors faire appel au déplacement de points de contrôle situés à la surface du visage. Le nombre de points de contrôle est de l'ordre de 35 dans [Mer07] et 70 dans [SK08].
- La direction du regard peut être mesurée à l'aide de deux angles.

[2]Toutefois, d'autres modélisations utilisées dans le domaine de l'animation 3D de signeurs virtuels peuvent faire appel à des squelettes sur lequel s'appuie la peau du visage. L'expression du visage peut donc être aussi représentée par un vecteur d'angles.

On le voit donc à partir de l'énumération qui précède, l'espace caractérisant l'état d'un signeur à un instant donné est de grande dimension. Par ailleurs, les données caractérisant cet état sont de nature très différente. Bien que la majorité puisse être décrite par un vecteur d'angle, chaque mesure n'a pas la même signification, ni la même importance relative dans la caractérisation du mouvement.
La précision de la caractérisation de l'état courant du signeur varie en fonction de l'appareillage de capture utilisé (cf. §4). Si les systèmes de capture de mouvements optiques ou mécaniques permettent de mesurer avec précision les différents angles relatifs aux postures et aux configurations manuelles, le traitement de vidéo ne permet qu'un accès partiel à ces données. Ceci est dû autant au fait que le problème de reconstruction 3D de posture est mal posé, qu'aux méthodes de traitement d'images utilisées qui ne permettent pas encore de mesurer précisément les paramètres relatifs aux configurations manuelles.
Une difficulté supplémentaire pour l'exploitation du signal est le fait que la perception des mouvement fait appel à la fois à la compréhension de la posture à un instant donné, à sa variation (vitesse) et à sa dérivée seconde (accélération).
Il se pose donc un problème fondamental de représentation du signal à traiter, qu'on peut rapprocher du problème de sélection des bandes de fréquences significatives à prendre en compte dans le cadre du Traitement Automatique de la Parole.

5.2.2 Difficultés dues à la variabilité des signes

La seconde difficulté dont il est nécessaire de tenir compte est la variabilité des signes. Cette variabilité, mentionnée dans [Vibilis00] est extrêmement importante. Par contre, il n'est pas possible de considérer systématiquement, comme dans le traitement automatique des langues orales, que la variabilité des signes est une déviation par rapport à une forme de référence du signe. Il s'agit bien au contraire d'une variation signifiante qu'il est nécessaire de modéliser et de quantifier, car elle est interprétable comme une flexion du signe (cf. §2.3.1).
Les variations de signes peuvent toucher différents paramètres. L'énumération qui suit n'est pas forcément exhaustive, mais donne une idée des différentes variations qu'il faudra être en mesure de prendre en compte dans une modélisation de signe :

- Certains signes peuvent être effectués à une ou à deux mains.
- Il est possible d'effectuer simultanément plusieurs signes (parallélisme).
- La géométrie des signes peut changer en fonction de leur utilisation. Ainsi, le placement du signe, son orientation et son amplitude sont fréquemment amenés à être modifiés suivant la structure de l'énoncé, ou les nuances apportées au signe.
- La structure temporelle peut être amenée à changer entre deux utilisations d'un même signe. Ainsi, il est courant que des signes soient présentés par des signeurs comme possédant des

répétitions, mais que les répétitions disparaissent lorsque ces signes sont utilisés en contexte. D'autre part, le nombre de répétitions n'est pas forcément constant d'une utilisation à l'autre. Ce problème des répétitions est souligné en particulier par [SWP98].

– Les configurations manuelles utilisées pour la réalisation du signes peuvent être amenées à varier dans certains signes. Si on prend l'exemple du *verbe* [DONNER], la configuration de la main dépendra de l'objet que l'on donne.

La variabilité affecte donc l'ensemble des paramètres du signe ainsi que sa structure temporelle. Les modèles d'analyse devront prendre en compte cet aspect.

5.2.3 Difficultés dues à l'iconicité

La troisième difficulté est la nature iconique des LS. La présence d'un lexique non-standard rend d'emblée impossible un inventaire et une modélisation exhaustive de l'ensemble des unités gestuelles qui composeront les phrases à traiter. Pour autant, un modèle de traitement automatique des LS abouti devra être en mesure de modéliser ces unités et de traiter les énoncés qui les contiennent.
Un problème lié à l'iconicité et à la multilinéarité des LS est la présence d'unités non-manuelles porteuses de sens comme l'expression du visage, le balancement des épaules du signeur, l'orientation du buste et le regard. Ces unités interviennent peu au niveau lexical, mais ont une grande importance dans la structure de l'énoncé.

5.2.4 Difficultés dues à la grammaire spatiale

Les méthodes de reconnaissance de TAP les plus populaires sont essentiellement temporelles, autant au niveau lexical qu'au niveau de l'énoncé. Les grammaires utilisées dans le cadre du traitement automatique des langues vocales sont principalement basées sur des grammaires et des modèles de langages statistiques qui ne prennent pas en compte la dimension spatiale.
Cela pose doublement problème pour une application directe des méthode du TAP aux LS. D'une part, le fait que l'ordre des signes soit beaucoup moins contraint que dans les langues vocales rend plus difficile l'utilisation de modèles statistiques basés sur des enchaînements bigrammes ou trigrammes de signes. D'autre part, l'ignorance des paramètres spatiaux conduit à une reconstruction ambiguë du sens des énoncés.

5.3 Problème de représentation du signal

Nous venons de le voir, les problèmes d'adaptation des méthodes du traitement automatique de la parole sont multiples. Le premier d'entre eux est la quantité importante et l'hétérogénéité des données à traiter (cf. §5.2.1).
Il convient dès à présent de distinguer plusieurs cas de figure, suivant la méthode de capture de l'énoncé :

- Si le corpus est capturé *via* des dispositifs de capture de mouvements, il est possible d'avoir à chaque instant une quantité importante et précise de données sur la posture courante du signeur. A titre indicatif, un suivi par marqueur passif comme celui décrit dans la section §4.1.4.2 permet de mesurer à chaque instant la position 3D d'une centaine de marqueurs à la surface du signeur.
- Si le corpus est uniquement filmé par une seule caméra. On disposera aussi de nombreuses informations, parfois moins précises, comme la position 2D des différents membres du signeur, les déformations du visage, la forme de la main du signeur.
- L'acquisition par un système multi-vue constitue un cas intermédiaire qui permet de reconstruire partiellement l'information 3D.

Dans tous les cas, se pose le problème du choix des données à traiter pour une caractérisation ou une reconnaissance des signes. L'idéal serait, comme le propose [HCGM06], de dissocier les paramètres permettant d'identifier le signe des autres paramètres liés au style du signeur. Malheureusement, aucune étude n'a à ce jour proposé de méthode satisfaisante pour dissocier les gestèmes[3] du style du signeur.
Nous abordons dans les parties qui suivent différentes solutions pour parvenir à une représentation du signal plus facile à traiter automatiquement.

5.3.1 Compression du vecteur d'entrée

Une des premières solutions pour obtenir un signal plus synthétique est de le représenter dans un espace de plus petite dimension, ou de l'échantillonner avec une résolution temporelle inférieure. Ceci présente surtout un intérêt dans les cas où on dispose de *trop* d'informations, par exemple pour les données issues de capture de mouvement.

[HCGM06] montre que l'espace des posture peut être réduit par Analyse en Composante Principale

[3] gestèmes : composants de bases du geste

(ACP) à une dimension quatre [4]. Même dans cette nouvelle base, l'énoncé restitué par un signeur virtuel reste intelligible.
[FF05] propose une variante de cette approche en utilisant des ACP pondérés. La représentation des mouvements dans des bases parcimonieuses est d'autant plus intéressante que le nombre de composantes utilisées pour la reconstruction d'un mouvement peut être un indice intéressant pour une reconnaissance ultérieure de mouvements. Soulignons toutefois un inconvénient des méthodes d'ACP. Les composantes principales optimales pour la reconstruction d'un signal sont dépendantes du signal. Il faut donc s'assurer que le corpus qui servira à déterminer la nouvelle base sera bien représentatif, du point de vue de sa variabilité, de l'ensemble des signes qui seront projetés par la suite sur les différents vecteurs qui la composent.
La compression par projection dans une autre base peut également être utilisée pour la représentation des configurations manuelles. Ainsi, [CSW95] utilise l'Analyse en Composantes Discriminantes (ACD) pour coder les configurations manuelles et reconnaître 28 signes. [HSA95] caractérise les contours de la main par des Smart-Snakes, ce qui lui permet aussi d'avoir une représentation compacte de la main par un faible de nombre de points de contrôle.

Parallèlement à cette compression dans le domaine spatial, il est possible d'effectuer une autre compression dans le domaine temporel. [Vat08] propose de modéliser les trajectoires sous forme de spline, puis d'utiliser les points de contrôle pour effectuer la reconnaissance de primitives. Bien que cette dernière approche soit appliquée à la reconnaissance de mouvements de souris d'ordinateur, la démarche nous semble transposable aux LS dans la mesure où nous savons que les points de rebroussements jouent un rôle important dans la reconnaissance des signes. C'est d'ailleurs en suivant une méthode similaire que [SSA92] compresse une animation d'avatar en sélectionnant des postures clés quand la vitesse est minimale ou quand il y a un changement brusque de trajectoire.

5.3.2 Transformation des coordonnées

Si les données disponibles pour la reconnaissance sont moins nombreuses, l'étape de compression du signal n'est pas forcément nécessaire. Par contre, le problème de la base dans laquelle doivent être représentées les données demeure.
Dans [CBA+96], l'auteur propose de focaliser le traitement des mouvements des mains. Il compare les performances de reconnaissance de signes en représentant les trajectoires manuelles dans les espaces suivants représentés figure 5.4 :

- Les coordonnées cartésiennes (x, y, z) des mains en position absolue ou relative par rapport à la tête.

[4] Chaque posture est alors codée comme une interpolation entre 4 postures de base.

- Les vitesses des mains $(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt})$ par rapport à un repère absolu ou à la tête du signeur.
- Les coordonnées des mains (r, θ, z) par rapport à la tête du signeur.
- La vitesse des mains $(\frac{dr}{dt}, \frac{d\theta}{dt}, \frac{dz}{dt})$ exprimées dans un repère cylindrique centré sur la tête du signeur.
- D'autres variantes comme l'utilisation de coordonnées sphériques.

D'une manière générale, il ressort de cette étude que le taux de reconnaissance est meilleur en utilisant les vitesses des mains et en utilisant un repère relatif lié à la tête du signeur.
Vogler va plus loin dans [VM99a] en montrant la supériorité des indices globaux (lignes, plan) sur les indices locaux comme la vitesse et la position des mains. Il faut cependant garder à l'esprit que l'étude présentée ne portait que sur la reconnaissance de 22 signes.

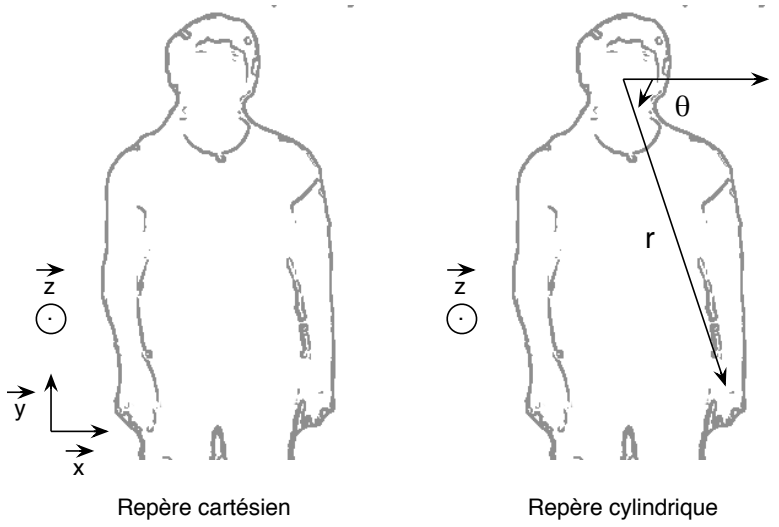


FIGURE 5.4 – Implantation des repères utilisés pour la reconnaissance

5.3.3 Solution composite

Nous venons d'énumérer plusieurs solutions pour représenter le signal d'une manière plus facilement utilisable dans le cadre du Traitement Automatique des Langues des Signes. Ces solutions de projection dans un autre espace de plus petite dimension, de rééchantillonage temporel et de changement d'espace de représentation ne sont pas mutuellement exclusives et sont amenées à être combinées dans de nombreuses études. On citera à titre d'exemple le travail présenté dans [BWK+04]. L'auteur dissocie dans un premier temps les paramètres de position absolue, de position relative des mains, de mouvement et de caractérisation des configurations. Ensuite, une projection est effectuée dans un espace de plus petite dimension par le biais d'Analyse en Composante Indépendante.

Une fois le signal convenablement représenté, les données peuvent être traitées par différentes méthodes pour reconnaître ou caractériser les signes. Dans les sections qui suivent, nous passons en revue les méthodes de Réseaux Neuronaux, de Dynamic Time Warping et de Modèles de Markov Cachés qui sont les plus utilisées dans le domaine, puis nous mettons en avant plusieurs méthodes alternatives.

5.4 Réseau neuronaux pour la reconnaissance de signes

Commençons par souligner le fait que les réseaux neuronaux sont plutôt adaptés à un traitement de données statiques qu'à un signal dynamique. Ceci explique pourquoi leur principal domaine d'application dans le domaine du TALS est la reconnaissance des configurations manuelles. Ainsi par exemple, [Mau09] utilise un perceptron pour la reconnaissance de configurations manuelles.
Toutefois, plusieurs études mentionnent également l'utilisation de réseaux neuronaux pour la reconnaissance de signes. [MHTAM07] traite dans un premier temps des images statiques à partir d'une transformée de Hough puis insère les résultats dans un réseau de neurones afin de reconnaître le signe qui aurait pu générer l'image. [WK95] utilise aussi des réseaux de neurones qui reçoit en entrée, une caractérisation du début et de la fin des signes à reconnaître. Il obtient 96% de reconnaissance sur un ensemble de 14 signes. Les réseaux de neurones sont aussi utilisés entre autre par [Su00] et [KJB96]. Pour pouvoir traiter les signes dans le domaine temporel, certaines adaptations des réseaux de neurones sont nécessaires. Ainsi, [MT91] utilise des réseaux de neurones récurrents tandis que [YAT02] met en oeuvre des réseaux de neurones à délai temporel.

Dans l'ensemble, les conclusions issues des résultats des traitements de signes par réseaux neuronaux sont assez proches de celles obtenues dans le cadre du Traitement Automatique de la Parole. Les corpus utilisés pour la validation sont en général de l'ordre de quelques dizaines de signes seulement pour des performances de plus de 90% d'identification correcte. Les performances chutent dès que le nombre de signes devient plus important. Le temps nécessaire pour la reconnaissance d'un signe est très conséquent. A titre indicatif, [HH98] rapporte un temps nécessaire de traitement de 10s pour arriver à reconnaître un signe parmi 15.
Il semble donc que l'utilisation de réseaux neuronaux pour la reconnaissance de signe soit relativement peu efficace et les performances obtenues laissent peu vraisemblable le passage possible à une échelle de plusieurs milliers de signes.

5.5 DTW pour la reconnaissance de signes

Les méthodes basées sur le Dynamic Time Warping permettent d'effectuer un alignement temporel et une comparaison de deux signes. L'avantage de ces méthodes est qu'elles ne requièrent qu'un

exemplaire du signe à reconnaître. Ce signe peut alors directement être utilisé comme modèle de référence.
[DP93] utilise les DTW pour discriminer 4 signes monomanuels. Il utilise la corrélation de l'image de la main du signeur avec un certain nombre d'images de référence pour arriver à identifier les configurations manuelles. Les méthodes basées sur le DTW ont aussi été utilisées avec succès par [HCGM06] et [FF05] pour l'alignement de signes. Plus récemment, [Alo06] a utilisé une méthode de DTW pour localiser les exemplaires d'un signe dans une vidéo en LS.
Les méthodes de DTW ont comme inconvénient la relative lourdeur du modèle constitué pour chaque signe. Le modèle est en effet constitué par une caractérisation de la posture du signeur à chaque instant pendant la production du signe. Une des conséquences est un temps de calcul important pour effectuer un alignement et une comparaison de signes. Pour résoudre ce problème, [Alo06] propose d'abandonner l'alignement de séquences temporelles si leur similarité est inférieure à un certain seuil (élagage). [SSA92] suggère un rééchantillonage temporel irrégulier du signe de référence préservant la dynamique du mouvement, ce qui permet de diminuer le volume du modèle, tout en restant compatible avec un traitement efficace de l'alignement temporel basé sur la programmation dynamique.

5.6 Les Modèles de Markov pour la reconnaissance de signes

L'exemple du travail de [SSA92] que nous venons de citer montre la préoccupation de réduire la taille du modèle à un minimum d'états. Les Modèles de Markov Cachés (MMC) s'inscrivent dans cette perspective. Comme dans le cas de la parole, les MMC ont été abondamment employés dans le domaine de la reconnaissance de signes. Leur utilisation est basée sur l'hypothèse forte que le signe est une succession de gestèmes.
Les travaux visant à une reconnaissance des signes basés sur les MCC utilisent aussi bien des données issues de traitement de vidéos monovues [SHJ94] que des données de capture de mouvement [YAT02]. Les performances rapportées par les auteurs varient en fonction du dispositif d'acquisition et de la taille du vocabulaire. Un autre facteur important mentionné par plusieurs études est la dégradation des performances de reconnaissances lorsque le signeur exécutant le signe à reconnaître est différent de celui qui a produit le corpus d'apprentissage. A titre indicatif, [ZK05] rapporte un taux de performance de 99,3% pour reconnaître 232 signes, qui chute à 44,1% dans la version multi-signeur de l'algorithme. [AG98] fait état d'un taux de reconnaissance de 94% pour un corpus de 262 signes qui chute à 73% lorsqu'un signeur différent est utilisé dans les étapes d'apprentissage et de reconnaissance.
Plutôt que d'énumérer les nombreux travaux qui utilisent les Modèles de Markov Cachés, nous préférons souligner les adaptations nécessaires de ces méthodes au traitement des LS, en nous focalisant sur les problèmes de modélisation des phonèmes, de parallélisation des traitements, de la modélisation des transitions et de la prise en compte de la variation des signes.

5.6.1 La modélisation des gestèmes

Nous rappelions dans le paragraphe précédent que la modélisation de signes sous forme de modèles de Markov était fondée sur l'hypothèse que le signe était une séquence d'états. Pour les langues vocales, ces états correspondent approximativement à des phonèmes ou des parties de phonèmes que l'ont peut dénombrer (de l'ordre de 37 en français). Qu'en est-il pour les LS ?
Comme le souligne Vogler dans [VM99b] , un état correspond au moins à un emplacement, une orientation et une configuration manuelle pour chaque main. Selon l'auteur, il y aurait en tout plus de 10^{10} combinaisons possibles des différents paramètres.
Certaines études comme [WSG02] se sont engagées dans cette voie de modélisation de tous les phonèmes. Les modèles proposés contiennent en tout 2400 phonèmes. Les auteurs rapportent toutefois des résultats satisfaisants, uniquement lorsque les phrases à reconnaître appartiennent au corpus d'apprentissage.
Pourtant, nous savons par ailleurs qu'il existe un faible nombre de primitives de mouvement qui regroupent la majorité des signes [Bra96]. Les configurations impliquées dans la production des signes de la LSF peuvent être évaluées à un peu plus d'une cinquantaine [Bou09].
Il serait donc possible de diminuer le nombre de phonèmes à condition de traiter séparément les différentes composantes du signe (configuration, orientation, mouvement, labialisation, expression du visage ...).

5.6.2 Les Modèles de Markov Cachés et leurs variantes

Le traitement séparé des différentes composantes implique une modification de la topologie des Modèles de Markov Cachés. Ceci est le cas dans les modèles suivants illustrés dans la figure 5.5 :

- Les MMC couplés, où chaque état de l'instant t est lié à un état à l'instant $t + 1$. Ces modèles prennent en compte la synchronisation des différentes composantes du signe.
- Les MMC liés, où plusieurs états de l'instant t sont liés entre eux. Ces modèles prennent en compte l'interdépendance des paramètres à chaque instant de la production du signe.
- Les MMC parallèles, où les états sont synchronisés seulement en début et en fin de signe. Ces modèles prennent en compte uniquement le parallélisme des différents paramètres.

Une étude menée sur la reconnaissance de gestes de Taichi [BOP97] montre que les MMC couplés donnent des résultats meilleurs que les MMC liés en terme de reconnaissance. Vogler, dans [VM99a] [VSM00] utilise les MMC parallèles en utilisant deux chaînes correspondant aux deux mains. Les deux chaînes sont synchronisées en début et en fin de signe. L'auteur rapporte des résultats de 94,23% de succès sur 22 signes.

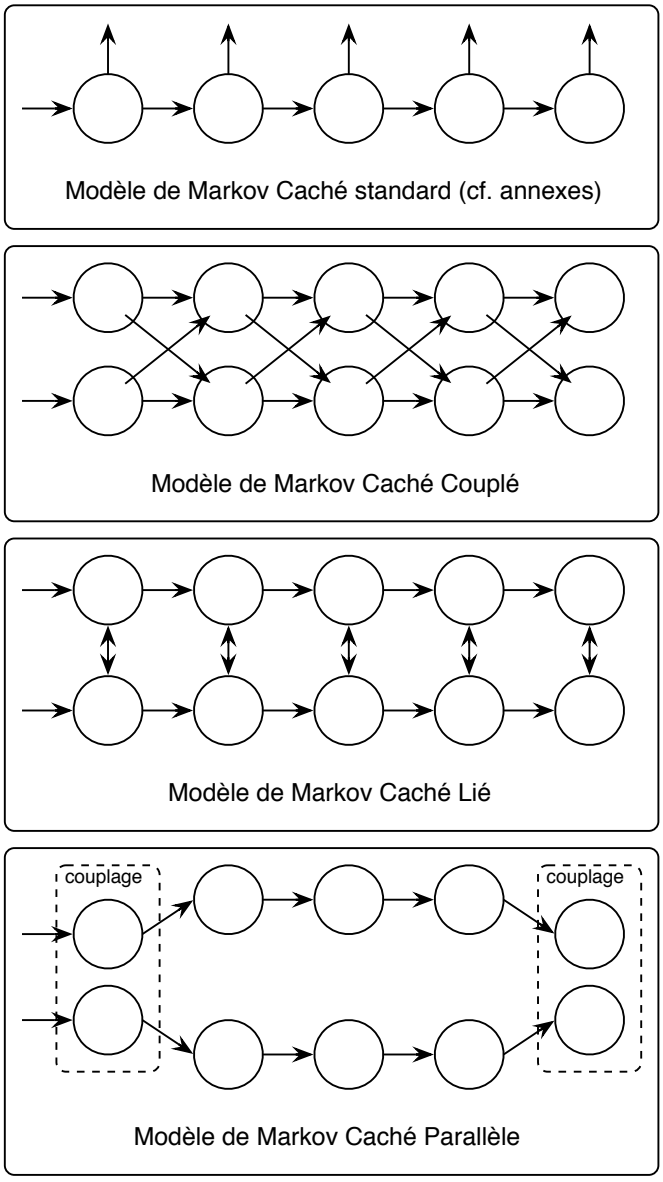


FIGURE 5.5 – Différents types de Modèles de Markov Cachés

Les Modèles de Markov Cachés Parallèles, en plus d'être économiques en terme de temps de calcul, présentent de nombreux intérêts par rapport aux MMC traditionnels :

- La parallélisation permet de réduire considérablement le nombre de gestèmes utilisés, et par conséquent, de réutiliser plusieurs fois une chaîne ou un état correspondant à un gestème. Par exemple, si le mouvement d'une main est modélisé séparément de sa configuration et de son orientation, il sera possible de réutiliser les mêmes Chaînes de Markov pour tous les signes impliquant des mouvements similaires.
- En LS, certains signes peuvent être effectués tantôt à une, tantôt à deux mains. De même, il est possible d'effectuer de temps en temps deux signes différents avec les deux mains. Le

parallélisme des Chaînes de Markov permet de prendre en compte de manière élégante ces deux phénomènes en utilisant des modèles de signes qui peuvent être adaptés séparément pour chaque main à condition que les signes commencent et finissent en même temps.

- Nous avons abordé uniquement le traitement des paramètres liés aux mouvements de la main. Il faut cependant savoir que les MMC parallèles peuvent également être utilisés pour traiter les labialisations, les expressions du visage, ou les mouvements de la tête.
- Il est possible d'ajuster le nombre d'états de chaque chaîne en fonction de la variabilité du paramètre qu'elle modélise.

Cette idée d'utiliser des modèles de primitives de mouvement élémentaire et de les combiner pour modéliser des signes plus élaborés, a été reprise par plusieurs auteurs. Parmi eux, [KC01] utilise des Chaînes de Markov Caché pour parvenir à reconnaître les mouvements balistiques, les arcs de cercle et les mouvements circulaires. Les primitives plus complexes comme les "mouvements en point d'interrogation" et les "répétitions de mouvement balistique" sont exprimées comme une séquence de mouvements élémentaires. Les résultats de reconnaissance montrent des taux de performance de 94% à 97% d'identification correcte des mouvements selon les conditions d'expérimentations.

D'autres variantes des Chaînes de Markov ont été utilisées avec plus ou moins de succès dans le cadre du traitement automatique des LS :

- Les MMC entrée-sortie (IOHMM) ont été testées par [MBVC00], mais [JM09] montre que les HMM ont des meilleures performances que les IOHMM.

- [KH97] utilise des Modèles de Markov Partiellement Cachés et reporte une amélioration de 73% sur la reconnaissance de 4 signes. La faible taille du corpus et le fait que cette approche n'ait pas été utilisée par d'autres équipes ne permet pas de conclure sur l'apport réel de cette *amélioration*.

De la même manière que [Alo06] pour les DTW, [FGZ04] met en oeuvre une technique d'élagage pour n'évaluer que les chaînes de Markov qui ont le plus de chance de correspondre à un signe.

5.6.3 Modélisation des transitions

En LS, les signes sont séparés par des transitions. Ces gestes permettent aux différents articulateurs de passer de la posture finale d'un signe à la posture de départ du signe suivant. Ce problème est relativement peu soulevé dans le cadre des Langues Vocales et est résolu par l'insertion de modèles

de silence et la modélisation de la coarticulation par le biais de modèles bigrammes ou trigrammes.
Le problème de la prise en compte des transitions dans les MMC pour les LS a été soulevé par [LO98]. Tandis que plupart des modèles de LS utilisés dans les Modèles de Markov Cachés modélisent les transitions comme un seul état, [VM97] tente de faire des modèles de coarticulation plus précis. L'auteur observe que le modèle à utiliser est en fait majoritairement dépendant des positions des mains au début et à la fin de la transition. Cette observation va dans le sens de l'utilisation d'un modèle de trigramme pour les transitions.

5.6.4 Prise en compte de la variabilité

Le plus grand problème auxquelles se heurtent les Modèles de Markov Cachés est la variabilité des signes à reconnaître. Cette variabilité signifiante est d'un autre ordre que celle d'une différence de prononciation qu'on pourrait observer dans les langues vocales.
Un autre type de variabilité est lié au signeur. Chaque signeur a une morphologie et un style dans la production des signes qui lui sont propres. Cela pose des problèmes lorsque le signe à reconnaître a été produit par un signeur qui ne fait pas partie de la base d'apprentissage.
Une variation (changement d'orientation, translation, changement d'amplitude) affecte en général la totalité du signe si bien qu'elle se répercute sur chaque état du modèle. Nous distinguerons dans les parties qui suivent les méthodes prenant en compte la variabilité liée à la structure spatiale des signes, à leur structure temporelle, et la manière d'inférer la variabilité d'un signe à partir de peu d'exemples.

5.6.4.1 Variabilité spatiale

La variabilité spatiale est, entre autre, une conséquence de la relocalisation des signes dans l'espace de signation. On assiste alors à une translation du signe dans un autre point de l'espace. Partant du constat que le même déplacement se répercute dans les différents états successifs du signe, [Alo06] propose de déduire la translation du signe à partir du premier état, et de propager cette translation dans les états suivants du signe. Cette méthode est par conséquent très sensible à des erreurs de mesure sur les premiers instants de la réalisation du signe.
La variabilité spatiale peut aussi être la conséquence de l'utilisation d'une main ou de deux pour l'effectuation même d'un signe. Pour résoudre ce problème, [Bos02] met en œuvre trois traitements parallèles correspondant aux hypothèses "signe effectué avec la main droite", "signe effectué avec la main gauche" et "signe effectué avec les deux mains". L'hypothèse la plus vraisemblable est ensuite retenue.
Lors des transferts personnels et situationnels, il peut arriver qu'un signe soit effectué dans différentes orientations. [NW96] propose de projeter la trajectoire des mains du signeurs sur le plan 2D parallèle aux trajectoires manuelles afin de les normaliser. L'auteur reporte un succès entre 96% et 100% dans la reconnaissance d'une dizaine de primitives de mouvement.

Les signes sont affectés par de nombreuses autres flexions spatiales comme les changements d'amplitude ou de direction. Pour les prendre en compte dans le modèle [WB99] exprime les états des Modèles des Markov Cachés en fonction d'un certain nombre de paramètres. L'algorithme de MMC paramétrique est évalué sur la détection de mouvements de pointage d'indication de dimension dans le cadre d'Interaction Homme Machine. Les auteurs soulignent une amélioration significative des performances de reconnaissance grâce à l'introduction de paramètres.

Nous le voyons au travers des exemples précédents, la variabilité des signes conduit les différentes équipes de chercheurs en traitement du geste à dégager les paramètres de production du signe, et à les normaliser de manière à rendre l'identification plus aisée. Si on regarde comment ces différents algorithmes sont mis en oeuvre, on s'aperçoit que le jeu de paramètres doit être estimé pour chaque intervalle temporel de la séquence à reconnaître. Pour une détermination optimale des paramètres, [WB99] utilise même une approche itérant les étapes d'alignement et d'estimation des paramètres rendant les calculs extrêmement coûteux.
Dans les deux cas, la programmation dynamique ne peut plus être utilisée avec autant de succès pour la reconnaissance de signe, ce qui remet en cause l'intérêt des Modèles de Markov Cachés.

5.6.4.2 Variabilité temporelle

La variabilité temporelle vient souvent s'ajouter à la variabilité spatiale que nous venons de mentionner, elle affecte à la fois la vitesse d'effectuation du signe et sa structure temporelle.
En ce qui concerne la variation de vitesse d'effectuation du signe ainsi que l'insertion de pauses en début en fin de signe, les MMC apportent déjà une solution puisque le modèle supporte n'importe quelle déformation temporelle préservant les relations d'antériorité entre les différents états.
En revanche, la topologie acyclique des Chaînes de Markov traditionnellement utilisée pour les LV, n'est pas compatible avec une variation du nombre de répétitions qui intervient justement fréquemment dans les signes des LS comportant des répétitions. Ce problème est bien souligné par [Alo06] qui suggère de créer une Chaîne de Markov différente pour chaque nombre de répétition du signe. Il suffit pourtant de modifier la structure des chaînes de Markov pour y inclure des cycles comme le fait [LPK$^+$97]. Une génération automatique de Modèles de Markov cycliques à partir d'exemples de signes est proposée dans [BWK$^+$04]. L'auteur parvient à identifier correctement 43 signes avec 97,6% de succès.

5.6.4.3 Générer de la variabilité à partir de peu d'exemples

L'entraînement des MMC requiert un nombre important d'exemples. L'apprentissage d'un nouveau modèle demande donc en général de gros volumes de corpus annotés qui peuvent être issus de re-

transcription écrite de documents audios pour les langues vocales. Nous ne disposons pas, pour les LS, d'important volumes de données transcrites dans le cadre du traitement automatique pour plusieurs raisons :

- Les LS n'ont pas de forme écrite, ce qui oblige les annotateurs à transcrire les corpus dans une autre langue (glose).
- Toutes les unités gestuelles n'ont pas forcément un équivalent sous forme d'un mot en langue vocale. Ceci est particulièrement vrai pour les unités gestuelles iconiques.
- La thématique de l'annotation d'un énoncé en LS est une problématique très récente, si bien que nous ne dispons pas encore d'un volume d'annotations important.
- La contrainte de respect du droit à l'image limite les bases de vidéo sur lesquelles on peut procéder à des annotations. Nous utilisons donc le plus souvent des corpus réalisés explicitement dans ce but.

Nous nous trouvons alors devant un problème de taille : celui de prédire la variabilité d'un signe à partir d'une, ou de peu de variantes de réalisation.
Plusieurs solutions sont proposées dans la littérature. L'utilisation de MMC parallèle en est une. Il est possible d'optimiser un Modèle de Markov représentant une primitive de mouvement à partir de plusieurs exemples de réalisation de ce mouvement. Ce Modèle de Markov pourra être ensuite réutilisé dans les modèles de tous les autres signes utilisant la même primitive, même si nous n'en connaissons qu'une seule réalisation.
Une autre solution est de décréter la variabilité des différents paramètres du signe *a priori*. [BW97] propose ainsi de modéliser un signe comme une succession d'états admettant une fonction d'appartenance floue.

5.7 Modélisations alternatives des signes

Les méthodes les plus populaires que sont les MMC et les DTW modélisent un signe comme un succession d'états. Avec les MMC parallèles, une étape est franchie en traitant de manière simultanée les différents paramètres sur des chaînes différentes. Nous verrons dans cette partie qu'il existe d'autres méthodes de caractérisations des paramètres que les Modèles de Markov, basés sur les spécificités des signes utilisés dans le LS. Nous aborderons successivement les problèmes de caractérisation des mouvements et de critères de segmentation.

5.7.1 Caractérisation des chérèmes

Analysées sous un angle linguistique, les méthodes que nous présentons dans cette partie font clairement référence à un modèle paramétrique de signe dans lequel chaque signe peut être représenté par

un ensemble de traits distinctifs que [SCC78] nommait chérèmes. Il s'agit alors de mettre en oeuvre la méthode appropriée pour arriver à distinguer les différents paramètres.
[DWT04] propose une comparaison de chérèmes, basée sur une prise en compte des mouvements horizontaux et verticaux de la main. L'auteur n'applique toutefois cette approche qu'aux signes monomanuels.
Cette démarche est aussi adoptée par [STO98] qui propose d'associer une valeur numérique à chaque chérème et de calculer la distance entre deux signes comme une distance de Mahalanobis dans un espace dont les dimensions sont les chérèmes.
Citons une étude isolée appliquée à la reconnaissance de gestes communicatifs et dont les résultats sont particulièrement intéressants pour le traitement automatique des LS. Dans [BE95], le but est de reconnaître des types de mouvement comme "ouvrir", "arrêter", "porter". L'algorithme de classification est basé sur la prise en compte d'indices tels que le travail contre la gravité, l'accélération, l'amplitude, l'écartement des mains, l'angle des avants bras, la proximité au corps et la répétition. Nous sommes très proche d'une caractérisation morphémique d'un signe. Cette approche pourrait permettre de préciser le sens d'unités gestuelles de grande iconicité non-standardisées dans le cadre d'un traitement automatique.

On peut aussi mentionner un certains nombre de méthodes spécialisées dans la détection de signes à répétition. Par exemple, [CJM04] utilise une méthode d'autocorrélation pour détecter les gestes périodiques. [PNN94] utilise uniquement l'information de variation d'intensité des différents pixels de l'image sans effectuer de suivi des différentes parties du corps et arrive à identifier différents mouvements périodiques tels que la course et la marche. [DWT04] met en place une méthode à part pour traiter les signes de LS comportant une répétition. Nous tenons cependant à souligner qu'il y a lieu d'être prudent sur l'adaptation des méthodes de traitement de geste à répétition aux LS. Contrairement aux gestes impliquant un grand nombre de répétitions comme la marche et la course, les signes admettent en général une ou deux répétitions qui ont le plus souvent des amplitudes différentes (cf. §7.3.2).

5.7.2 Séparation de la reconnaissance et de la segmentation

La plupart des approches basées sur les MCC opèrent simultanément une reconnaissance et une segmentation des signes. Toutefois, d'autres méthodes comme [LPK$^+$97] opèrent à une segmentation avant l'étape de reconnaissance.
Dans [KC01], un automate à états finis est utilisé pour parvenir à une segmentation des énoncés. Chacun des état est une primitive de mouvement comme les mouvements balistiques, les arcs de cercles ou les mouvements circulaires.
[VM98] emploie une méthode proche en décomposant les mouvements en phases de mouvements balistiques et de tenues, en caractérisant les mouvements par le plan dans lequel ils sont effectués.

Ceci permet aussi à l'auteur d'aboutir à une segmentation.
Il est aussi possible d'utiliser les passages des mains par une vitesse minimale pour opérer une pré-segmentation d'un énoncé. Toutefois, il est important de garder à l'esprit que les points de rebroussement où la vitesse des mains s'annule peuvent aussi être présents à l'intérieur de la réalisation d'un signe.
Ces approches sont intéressantes dans la mesure où elles impliquent qu'il serait possible d'opérer une segmentation d'un énoncé en signe à partir du mouvement seul, sans passer par une étape de reconnaissance.

5.8 De la reconnaissance de signes isolés à la reconnaissance en continu

Nous avons abordé jusqu'ici le problème de reconnaissance ou de caractérisation du signe isolé. Cette section est consacrée au problème de reconnaissance du signe dans un énoncé. Dans ce domaine, on note également la prédominance des méthodes basées sur les Modèles de Markov Cachés. Comme pour les langues vocales, les statistiques sur les enchaînements de signes sont utilisées pour créer un modèle de langue et trouver l'enchaînement de signe le plus probable correspondant à l'énoncé. Cette approche pose cependant plusieurs problèmes :

- Pour obtenir un modèle de langue réaliste, un volume important de corpus annoté est nécessaire. Nous ne disposons pas d'un tel corpus pour les raisons mentionnées au §5.6.4.3.
- Un modèle de langue statistique est en général uniquement temporel. Dans les LS, l'ordre des signes est en général beaucoup moins strict que pour les langues vocales. De plus, la syntaxe des LS fait appel à la spatialisation qui est ignorée par les modèles de langue statistiques traditionnels.
- L'ensemble de signes utilisable n'est pas dénombrable car des unités gestuelles de grande iconicité peuvent être créées à partir de proformes de base[5]. Un modèle de langue prenant en compte les probabilité d'enchaînement entre tous les signes n'est donc pas envisageable.
- Du fait du parallélisme, la structure des énoncés des LS n'est pas uniquement séquentielle.

Malgré tout, les modèles de langue statistiques ont été utilisés dans le cadre du traitement automatique des LS dans des études comme [SP95]. [VM97] utilise également cette approche pour reconnaître une quarantaine de signes en contexte. Cependant, ces deux dernières études ne portaient que sur la reconnaissance de signes standards.

[5] [Cux00] dénombre une quarantaine de proformes qui peuvent être combinés dans des unités de grande iconicité.

La reconnaissance de signes composant un énoncé ne peut pas être limitée à une simple glose. Des signes comme les pointages doivent être davantage caractérisés pour pouvoir rendre l'énoncé qui les contient interprétable [ST00]. Les travaux de TALS intégrant cette spatialisation sont encore peu nombreux. Nous n'en citons que quelques uns dans les lignes qui suivent car la thématique d'interprétation de signes en contexte se trouve en marge des travaux présenté dans cette thèse. Dans [Bra96], l'auteur montre l'architecture d'un vrai système prenant en compte la dimension spatiale et l'évalue sur un petit ensemble de signes. [ST99] s'attache à l'interprétation de verbes directionnels et à leur flexion pour les accorder dans l'espace. Espérons que les modèles informatiques de LS comme celui développé dans [DL05] permettront d'intégrer mieux la spatialisation des LS dans les futurs travaux de Traitements Automatiques des Langues des Signes.

En plus de la spatialisation, il sera également nécessaire de mettre l'accent sur la détection et le suivi des paramètres non-manuels. Comme le font remarquer les auteurs de [OR05], la prise en compte de ces indices est encore très limitée, voire inexistante dans la plupart des études. Citons à titre d'exception l'étude [MGW00] qui prend en compte la labialisation en complément des paramètres manuels pour améliorer les performances de reconnaissance. Dans ce travail, la synchronisation des paramètres manuels et non manuels est effectuée à l'aide de MMC Parallèles. L'étude de [Par03] présente aussi une méthode d'interprétation des mouvements de la tête en complément de celui des mains pour arriver à reconnaître les négations.

Enfin, dès lors qu'il ne s'agit plus de reconnaître des signes isolés mais un énoncé, on peut faire intervenir des unités sémantique ou pragmatique (dialogue) de plus haut niveau, et exporter des données contextuelles. On est alors dans le cadre de l'étude d'un système dans lequel les traitements ne sont plus seulement ascendants (du signal vers l'interprétation), mais aussi descendant et donc prédictifs. Ceci permet d'introduire des contraintes supplémentaires dans les méthodes que nous avons présentées, en permettant de lever l'ambiguïté, voire de développer de nouvelles méthodes (de type vérification de prédiction). Ceci ne rentre pas dans le cadre de cette thèse.

5.9 Bilan

Les approches de traitement automatiques des LS sont multiples mais on note une claire prédominance des méthodes basées sur les Modèles de Markov Cachés. Le choix de ces outils peut être expliqué par plusieurs facteurs comme le nombre important d'algorithmes déjà disponibles, le cadre mathématique rigoureux des MMC et les performances obtenues dans le cadre du traitement automatique des langues vocales.
Des modification de la structure traditionnelle des chaînes de Markov sont nécessaires pour prendre en compte le parallélisme, la variabilité et la structure temporelle des signes. Les adaptations apportées par les différentes études auxquelles nous faisons référence portent sur les aspects suivants :

- Une représentation du signal favorisant une exploitation de phénomènes s'étendant sur des intervalles de temps de plus en plus long (détection de meilleur plan, d'amplitude d'un mouvement).
- Un traitement séparé des différents paramètres constitutifs du signe dans les MMC Parallèles .
- Une modélisation explicite de la variabilité des signes affectant l'ensemble des états.
- Une modification de la topologie des chaînes de Markov pour faire ressortir la synchronisation des différents paramètres et prendre en compte les signes à répétition.

Les performances obtenues à l'aide de ces méthodes sont assez difficile à comparer car les conditions d'évaluation diffèrent suivant les études. Cela est d'autant plus problématique que les performances varient en fonction de nombreux facteurs :

- Les performances chutent si le signeur ne fait pas partie de la base d'apprentissage [ZK05].
- Les résultats de reconnaissances sont moins bons si le signe est effectué en contexte.
- Les signes ne sont que rarement effectués par des signeurs ayant une LS comme première langue [OR05] lors des apprentissages et des évaluations. Derechef, lorsque le vocabulaire à reconnaître est très limité, les énoncés utilisés pour les textes sont tellement contraints que les signes sont utilisés avec moins de variabilité.
- Les conditions d'acquisition du corpus (fond fixe ou uniforme, mains nues ou colorées ...) influencent également les performances du suivi, et donc du système de reconnaissance.

Les meilleures performances de reconnaissance de signes dans une vidéo sont de l'ordre de quelques centaines de signes reconnus avec plus de 90% de succès. Les résultats sont en revanche bien meilleurs pour la reconnaissance basée sur des données issues de capture de mouvement. A titre indicatif, [FGZ04] reporte un taux de succès de 83,7% sur 5113 signes.
On voit également émerger en marge des approches traditionnelles, d'autres méthodes de reconnaissance ou de caractérisation de signes prenant davantage en compte les spécificités des LS. Ces approches fondées linguistiquement sur des modèles paramétriques de signe prennent plus en compte la variabilité des signes, tant au niveau des flexion spatiales qu'au niveau de la structure temporelle.
Les énoncés utilisés pour l'évaluation des méthodes sont souvent peu représentatifs dans leur composition, de la diversité des structures observables en LS. En particulier, il est rarement fait mention de l'utilisation de paramètres non manuels ou de la prise en compte des unités gestuelles iconiques non standard.
Pour permettre une meilleur prise en compte de la richesse des LS, nous prendrons pour constituer nos modèles, des productions de sourds dont la LS est la première langue. La plupart de ces énoncés n'ont pas été constitués spécifiquement dans un but d'analyse automatique de vidéo. Ceci ajoute une

difficulté pour le suivi du signeur dans la vidéo mais garantit une application de nos algorithmes de traitement dans une plus large palette d'applications. Le problème de suivi du signeur dans une vidéo sera d'ailleurs l'objet du prochain chapitre.

Troisième partie

Un modèle paramétrique au service du traitement automatique des Langues des Signes

CHAPITRE 6

SUIVI D'UN SIGNEUR DANS UNE VIDÉO MONO-VUE

6.1 Introduction

Dès lors qu'on souhaite effectuer une caractérisation objective des mouvements des signes produits par un signeur, une étape de suivi des différents membres sur la vidéo est incontournable. Le travail de recherche que nous présentons dans ce chapitre a permis d'aboutir à un suivi à la fois rapide et robuste de la tête, des mains, des coudes et du buste du signeur.
Nous exposerons dans un premier temps, des statistiques sur les différentes postures du signeur impliquées dans une production en LS. Les résultats issus de ces statistiques nous permettront de formuler des hypothèses de travail que nous utiliserons pour le suivi. Dans un second temps, nous présenterons la structure globale de notre algorithme de suivi, et nous détaillerons le traitement bas niveau de la vidéo permettant de passer d'une séquence d'images à la séquence de vecteurs de posture correspondants.

6.2 Caractérisation des postures d'un énoncé en LS

Notre but est de réaliser un algorithme de suivi d'un signeur vu de face. Il ne s'agit donc pas de concevoir un programme permettant de résoudre le cas général de reconstruction de la posture d'une personne filmée en monovue d'un point de vue quelconque. Le cadre d'application restreint que nous nous imposons permet d'utiliser à l'intérieur de l'algorithme de suivi (même au bas niveau), des spécificités de la LSF[1].
Nous présentons, dans cette partie, une série de statistiques réalisées sur des vidéos en LSF, permettant de caractériser la posture et de formuler des hypothèses que nous réutiliserons pour le suivi. Nous en proposons également une interprétation linguistique.

6.2.1 Corpus utilisé

Les mesures ont été effectuées sur une signeuse sourde. Le corpus utilisé provient du projet SignCom[2]. Les dix minutes d'énoncés que nous avons analysées portent sur la confection de cocktails, de salades et de galettes. Les productions sont à la fois des narrations et des dialogues. Nous avons choisi ce corpus car il implique énormément de transferts personnels et situationnels que nous savons

[1] On gagne ainsi en rapidité et en robustesse, ce que l'on perd en généralité.
[2] Projet ANR impliquant les laboratoires IRIT, IRISA, VALORIA, M2S portant sur la communication en LS entre des agents réels et virtuels.

FIGURE 6.1 – Cadrage de la vidéo du signeur

être riches en rotations du buste et de la tête.
Le corpus a été acquis simultanément par une caméra vidéo située en face du signeur et par des dispositifs de capture de mouvement à marqueurs passifs (cf. §4.1.4.2) de type VICON. Les vidéos sont filmées avec une résolution 640 x 480. Le cadrage des vidéos est semblable à celui présenté figure 6.1. Le système VICON utilise 12 caméras infrarouge. Trois d'entre elles sont situées à l'arrière, tandis que les autres sont positionnées devant le signeur à différentes auteurs. Le signeur est recouvert de plus d'une cinquantaine de marqueurs réfléchissants situés au voisinage des différentes articulations du signeur. Ceux-ci permettent à la fois de capturer la posture globale du signeur, ses configurations manuelles ainsi que ses expressions du visage. L'ensemble des mesures que nous exploitons sont issues de la reconstruction 3D de la posture du signeur déduite à partir de fichiers de capture de mouvements au format BVH (BioVision Hierarchy).

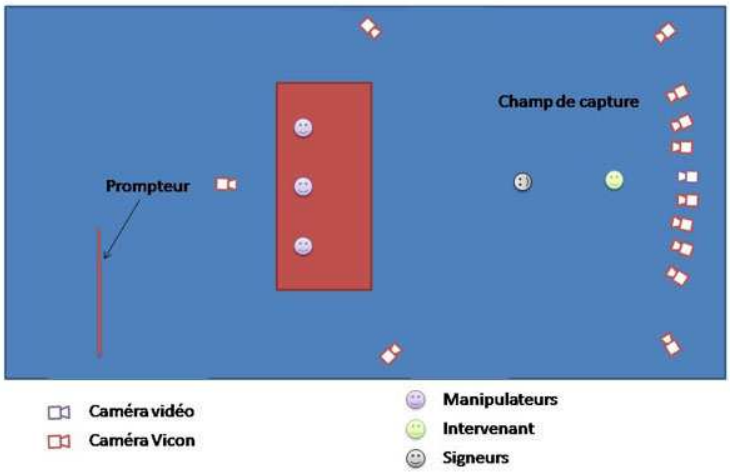


FIGURE 6.2 – Aménagement de la salle de capture

Lors de l'enregistrement du corpus, le signeur filmé s'adressait à un interlocuteur, lui aussi signeur, situé à côté de la caméra vidéo. L'interlocuteur sourd a la possibilité de visualiser une présentation visuelle (visible en haut à gauche de la figure 6.1) qui représente de manière iconique le contenu du corpus. Ce protocole de production permet d'une part d'éviter l'influence du français, et d'autre part de conserver un aspect naturel dans la production signée.

6.2.2 Rotation du buste

Certains algorithmes de suivis comme [BWK+04] utilisent l'hypothèse d'une invariance (ou d'une faible variabilité) de l'apparence du buste. Nous avons voulu vérifier jusqu'à quel point cette hypothèse était acceptable car nous savons d'après les recherches linguistiques que les rotations du buste et de la tête sont impliquées dans des phénomènes grammaticaux comme les transferts. Nous présentons donc ici une série de mesures sur les rotations du buste mesurées par le dispositif VICON.

6.2.2.1 Protocole

La mesure de l'orientation du buste dépend du point où elle est mesurée. Nous choisissons donc de la mesurer au point où les amplitudes des rotations sont maximales, c'est-à-dire au niveau des épaules. Les articulations utilisées pour l'évaluation des rotations du buste sont l'épaule droite Sr, l'épaule gauche Sl ainsi qu'une vertèbre au milieu du dos B. Nous utilisons également dans nos calculs le point Sm situé au milieu des épaules. La figure 6.3 illustre le repérage des différentes articulations.

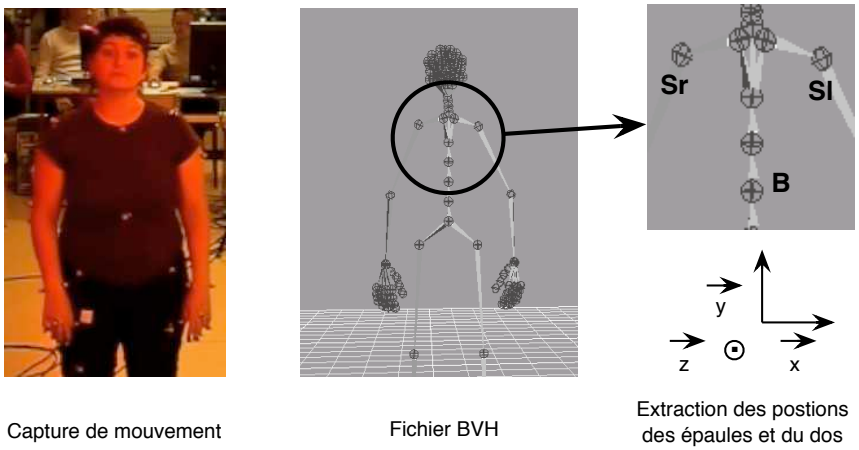


FIGURE 6.3 – Position des articulations utilisées pour calculer les rotations du buste d'un signeur

Nous choisissons d'évaluer séparément les rotations des épaules dans chaque plan : sagittal (θ_x), horizontal (θ_y) et frontal (θ_z). Ces différents angles sont représentés sur la figure 6.4 et sont calculés de la manière suivante :

$$\theta_y = atan\left(\frac{Sl_z - Sr_z}{Sl_x - Sr_x}\right)$$

$$\theta_x = acos\left(\frac{Sm_y - B_y}{BSm}\right) * sign\left(Sm_z - B_z\right)$$

On nomme $\vec{y'}$ la projection du vecteur $\vec{y}$ sur le plan (B, Sl, Sr).

$$\theta_z = \left(\vec{y'}, \overrightarrow{BSm}\right)$$

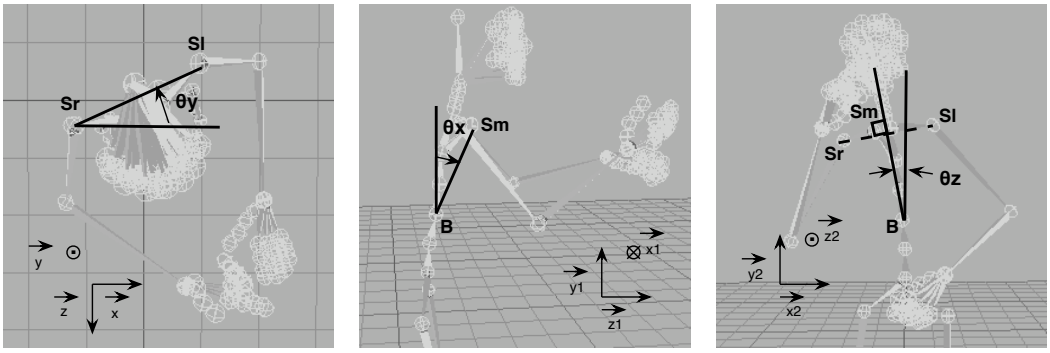


FIGURE 6.4 – Implantation des 3 repères $(O, \vec{x}, \vec{y}, \vec{z})$, $(O, \vec{x_1}, \vec{y_1}, \vec{z_1})$, $(O, \vec{x_2}, \vec{y_2}, \vec{z_2})$ par rapport aux épaules et au dos du signeur.

6.2.2.2 Résultats

Les résultats obtenus sont présentés dans les figure 6.5, 6.6 et 6.7. Les différentes orientations sont présentées sous forme de distributions.

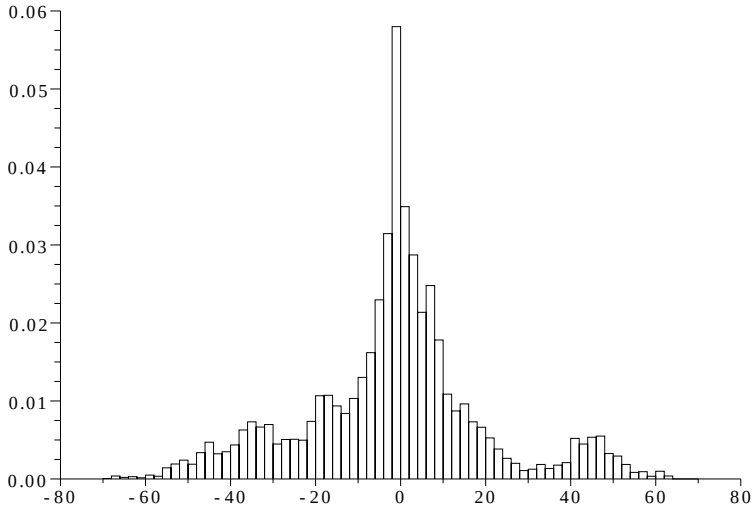


FIGURE 6.5 – Distribution de θ_y en degré

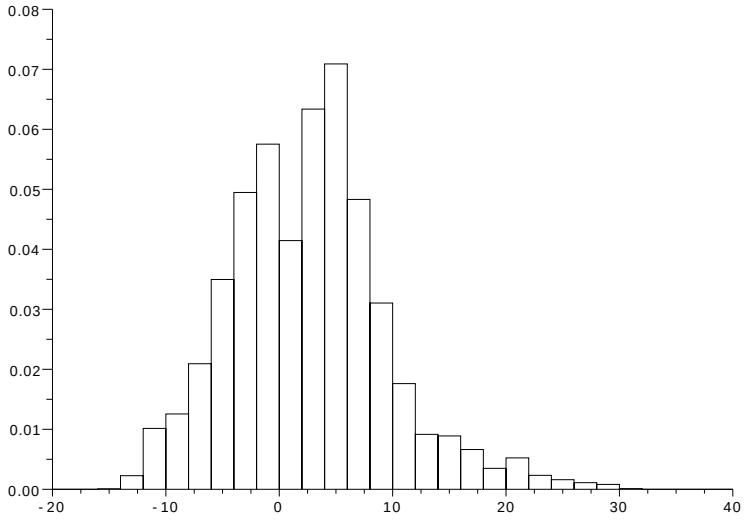


FIGURE 6.6 – Distribution de θ_x en degré

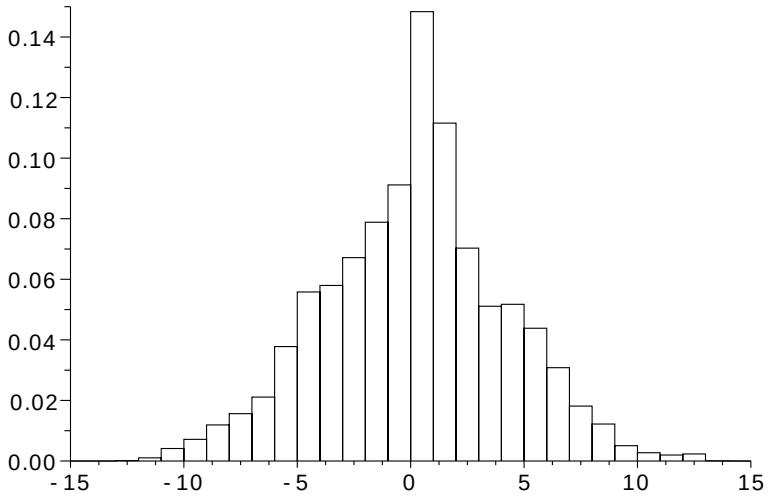


FIGURE 6.7 – Distribution de θ_z en degré

6.2.2.3 Interprétation

Nous formulerons des interprétations séparées pour chaque paramètre de rotation en soulignant les interprétations linguistiques éventuelles des résultats, ainsi que les conséquences pour le suivi.

La distribution de la rotation du buste dans le plan horizontal θ_y est centrée en -1.6° et admet un écart type de 21.4°. Le léger décentrage de la distribution est vraisemblablement dû au fait que le signeur était situé un peu de biais par rapport à l'interlocuteur.
La rotation du buste dans le plan horizontal engendre de faible variations de distance apparente entre les deux épaules dans la vidéo. L'écart type de ces variations est de 12% de l'écart moyen entre les deux épaules. Cet écart est significatif, mais compatible avec l'hypothèse d'une apparence du buste peu variable.

La distribution de la rotation du buste dans le plan sagittal θ_x est centrée en 2.6° et admet un écart type de 6.9°. Ce balancement du corps vers l'avant ou vers l'arrière est de faible amplitude et n'a que peu de conséquences sur la silhouette du buste vue de face.

La distribution de la rotation du buste dans le plan frontal θ_z est centrée en 0.2° et admet un écart type de 4.0°. Cette rotation ne change que légèrement l'apparence de la silhouette du buste. Par contre, θ_z cause un déplacement de la tête vers la droite ou vers la gauche qui peut facilement être estimé dans le suivi 2D.

Les analyse qui précèdent justifient l'hypothèse d'un modèle de buste fixe pour la détermination de postures dans le cadre de vidéos en LSF. Il est cependant important de souligner que ces rotations sont importantes dans la compréhension d'un énoncé en LS. Nous émettons l'hypothèse que les rotations du buste pourront être déduites en utilisant les positions de la tête, du buste et des coudes ainsi que les rotations de la tête qui ont une amplitude plus significative (cf. 6.2.3).

6.2.3 Rotation de la tête

De la même manière que précédemment, nous effectuons des statistiques sur l'orientation de la tête, de manière à caractériser l'amplitude des rotations de la tête dans chaque direction. Le protocole utilisé est identique à celui utilisé dans la caractérisation des rotations du buste. Nous nommerons τ_y, τ_x et τ_z les rotations dans les plans horizontal, sagittal et frontal.
Les points utilisés pour le calcul des rotations sont la base du crâne (à la jonction avec la dernière vertèbre), le nez du signeur ainsi que le sommet de la tête. La démarche retenue pour le calcul des angles est similaire à celle adoptée pour le calcul des rotations du buste.

6.2.3.1 Résultats

De la même manière que précédemment, les résultats obtenus sont présentés dans les figures 6.9 et 6.8. Les différentes orientations sont présentées sous forme de distributions.

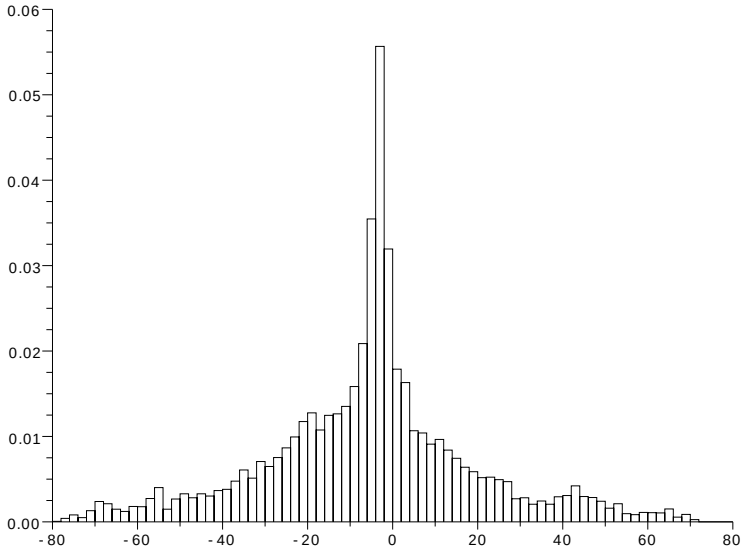


FIGURE 6.8 – Distribution de τ_y en degré

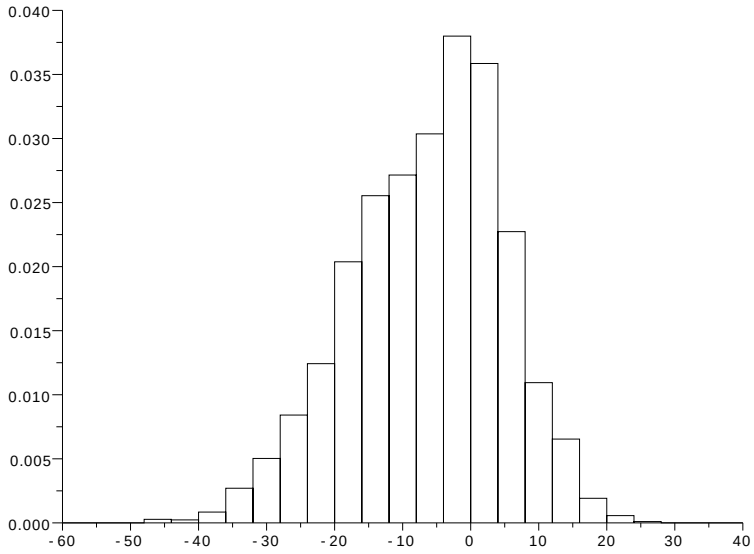


FIGURE 6.9 – Distribution de τ_x en degré

6.2.3.2 Interprétation

La distribution de la rotation de la tête dans le plan horizontal est centrée en -5.5° et admet un écart type de 24,9° légèrement plus important que l'écart type de la rotation du buste dans le même plan. Le décentrage de la distribution peut être expliquée par la position décalée de l'interlocuteur sourd par rapport au repère. La corrélation entre τ_y et θ_y est de 0.79 ce qui permet d'étayer l'hypothèse d'une détermination possible de l'orientation du buste grâce à la tête.
La distribution de la rotation τ_x de la tête dans le plan sagittal est centrée en -6.29° et admet un écart type de 11.2° beaucoup plus important que θ_x. De manière assez surprenante, on ne note pas de corrélation significative entre τ_x et θ_x (seulement 0.1). Il serait donc nécessaire de trouver d'autres indices dans l'image pour déterminer le balancement avant-arrière du buste.

Les statistiques qui précèdent montrent donc que les rotations de la tête ont une amplitude non négligeable. Ces rotations affectent naturellement l'apparence de la tête en projection dans la vidéo, mais n'occasionnent qu'une faible modification de la forme de la zone de peau de la tête. Nous choisissons par conséquent une modélisation grossière de la tête sous la forme d'une zone de peau rectangulaire.

6.2.4 Positions des coudes

Les données de capture de mouvement nous permettent également de connaître avec précision la position des coudes. Nous calculons à chaque instant :

- La position des coudes par rapport au buste dans le plan de la vidéo (la position du buste est prise arbitrairement comme le milieu des deux épaules),

- La profondeur du coude par rapport à la profondeur de l'épaule correspondante.

6.2.4.1 Résultats

Pour plus de lisibilité, nous superposons le diagramme représentant la distribution de positions du coude avec la silhouette de la signeuse.

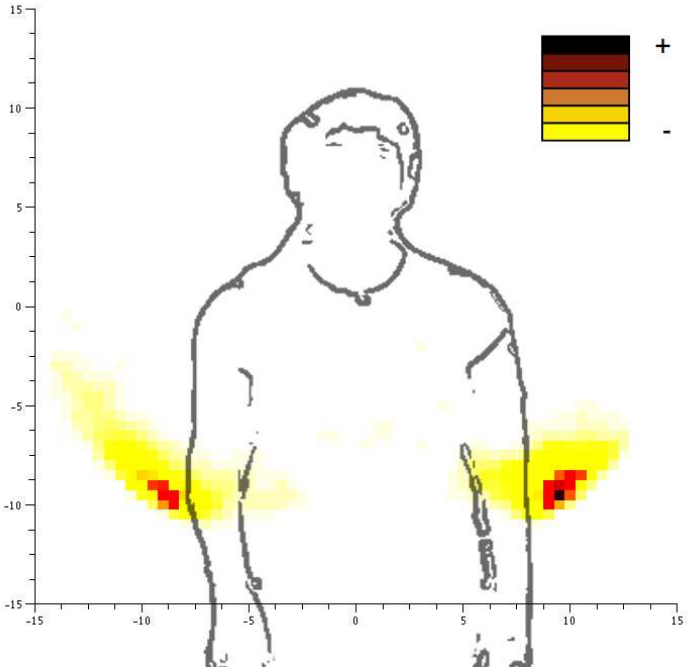


FIGURE 6.10 – Distribution de la position des coudes droite et gauche par rapport au buste

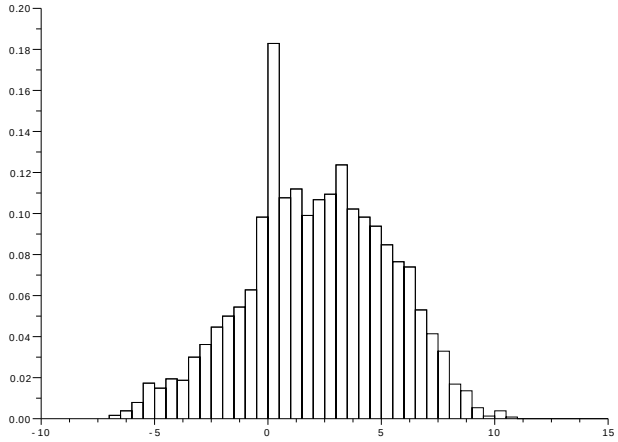


FIGURE 6.11 – Différence de profondeur entre le coude droit et l'épaule droite par rapport au repère $(O, \vec{x_1}, \vec{y_1}, \vec{z_1})$ (cf. fig. 6.4)

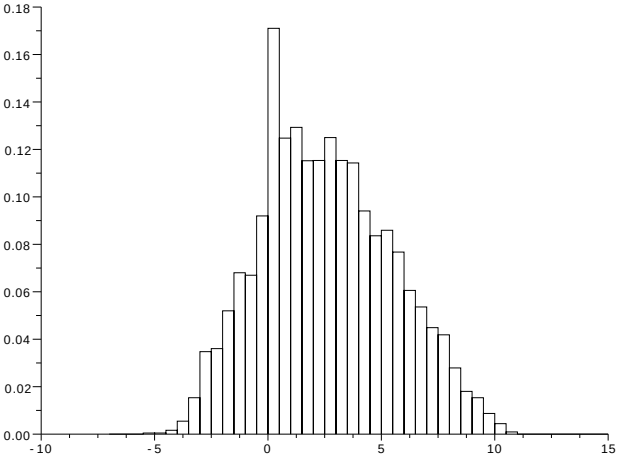


FIGURE 6.12 – Différence de profondeur entre le coude droit et l'épaule droite par rapport au repère $(O, \vec{x_1}, \vec{y_1}, \vec{z_1})$ (cf. fig. 6.4)

6.2.4.2 Interprétation

Plusieurs remarques s'imposent à partir de ces observations :

– La position des coudes en projection dans le plan $(O, \vec{x}, \vec{y})$ est relativement peu variable. La plupart du temps, ceux-cis sont situés à l'aplomb des épaules légèrement vers l'extérieur du

buste. Il sera donc nécessaire de trouver des détecteurs de traitement d'images rapides et efficaces qui fonctionnent correctement dans ce cas.

– Les positions occupées par les coudes sont limitées dans l'espace. Ceci découle directement de la contrainte cinématique liée à la longueur fixe du bras.

– Il existe certaines postures dans lesquelles le coude occulte le buste du signeur. Même si ce cas est peu fréquent, il faudra pouvoir le traiter dans l'algorithme de suivi.

– Les coudes sont, la majorité du temps, situés devant les épaules correspondantes ($Z_{coude} > Z_{épaule}$ dans le repère $(O, \vec{x}, \vec{y}, \vec{z})$ (cf. fig. 6.4)). Cette hypothèse est d'autant plus vérifiée que la profondeur est évaluée par rapport au plan vertical $(S_m, \vec{x_1}, \vec{y_1})$contenant les épaules du signeur.
Comme nous l'avons mentionné dans le §4.2.8, le problème de reconstruction monovue de la posture est un problème mal posé qui admet plusieurs solutions. Compte tenu des observations qui précèdent, nous émettrons l'hypothèse dans notre travail que les coudes sont situées à l'avant des épaules.

6.2.5 Positions des mains

Enfin, nous cherchons à caractériser la position des mains du signeur. Contrairement aux coudes et aux épaules, les mains ne sont pas des articulations. Parler de position des mains est donc en soi, presque un abus de langage. La position de la main sera considérée comme étant le milieu du segment reliant le poignet à l'articulation de la première phalange du majeur avec la paume.

6.2.5.1 Résultats

Les résultats sont exposés figure 6.13 où nous avons superposé la silhouette de la signeuse avec l'emplacement de ses mains. Les figures 6.14 et 6.15 montrent les positions relatives des mains en abscisse et en ordonnée. L'histogramme 6.16 montre la profondeur de la main droite par rapport à celle du coude droit.

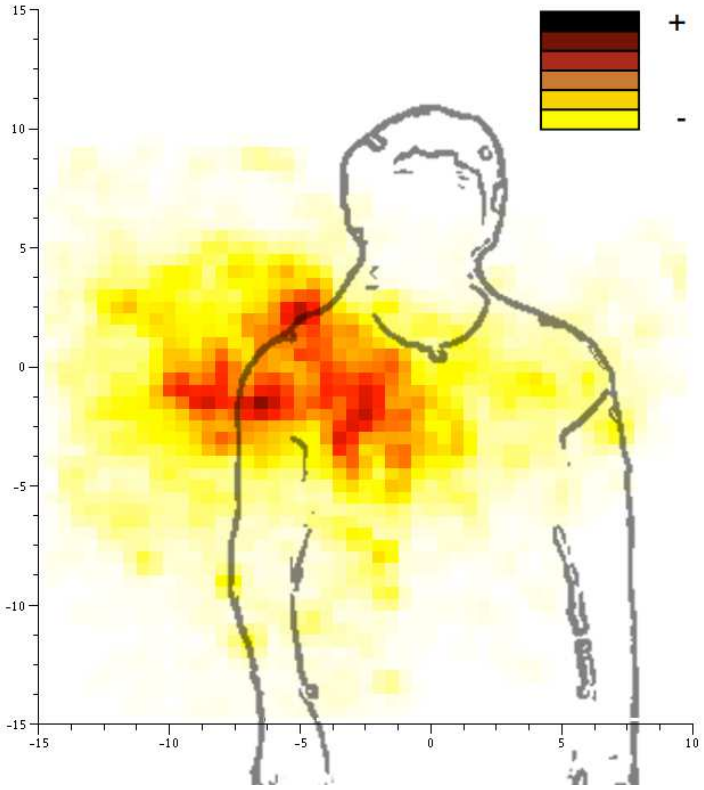


FIGURE 6.13 – Position du centre de gravité de la main droite de la signeuse

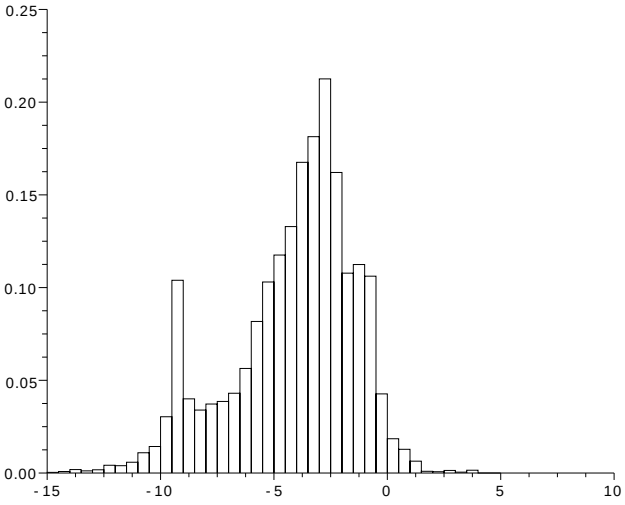


FIGURE 6.14 – Abscisses relatives des mains droites et gauches

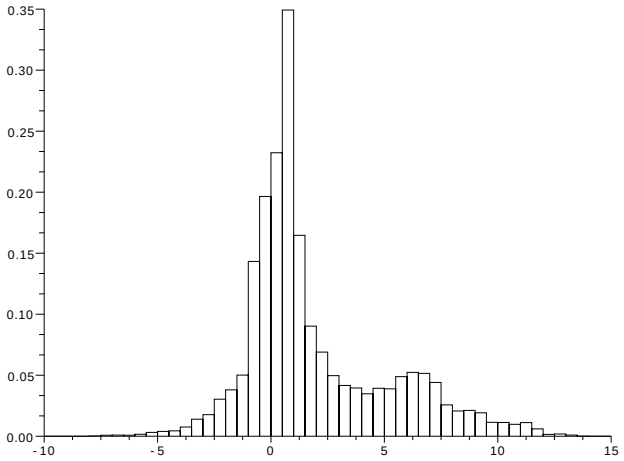


FIGURE 6.15 – Ordonnées relatives des mains droite et gauche

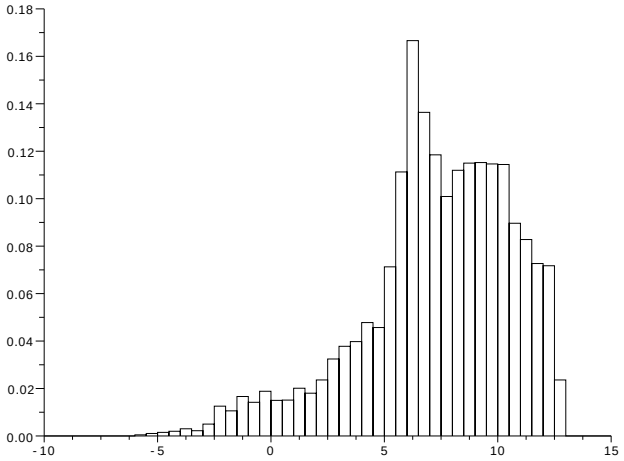


FIGURE 6.16 – Différence de profondeur entre la main droite et le coude droit

6.2.5.2 Interprétation

La main droite se trouve fréquemment devant le buste et la tête. Le graphique 6.13 nous indique qu'il sera nécessaire de gérer les occultations entre la main droite et la tête.

L'étude des positions relatives des mains droite et gauche du signeur révèle sans surprise que la main droite se situe le plus souvent à droite de la main gauche (du point de vue du signeur). Il est important à ce point de l'analyse, de mentionner le fait que la signeuse est droitière et utilise la main droite

comme main dominante, et qu'elle utilise par conséquent préférentiellement sa main droite pour effectuer les signes monomanuels. Ceci explique que la main droite soit la plupart du temps située au dessus de la main gauche de la signeuse. Cette asymétrie due à la latéralité du signeur peut être mise à profit pour désambiguïser les deux mains à partir de leur position. Nous utiliserons cette propriété au §6.6.1.

La comparaison des profondeurs du coude et de la main droite fait ressortir le fait que la main du signeur est majoritairement située devant le coude ($Z_{coude} > Z_{épaule}$ dans le repère $(O, \vec{x}, \vec{y}, \vec{z})$ (cf. fig. 6.4)).

6.2.6 Bilan

Ces différentes statistiques permettent de formuler une série d'hypothèses que nous utiliserons dans le cadre de notre algorithme de suivi :

1. Le buste a une apparence relativement peu variable. Il est donc possible de le modéliser sous la forme d'une silhouette indéformable. Il est cependant important de garder à l'esprit qu'il peut être nécessaire de conserver dans le modèle cinématique du corps du signeur, des degrés de mobilité correspondant aux différentes rotations du buste qui ont une signification dans le cadre des transferts.

2. La tête du signeur a également une mobilité relativement faible (de l'ordre de 15 à 25 degrés en moyenne dans chaque direction). Il est donc possible de la modéliser sous forme d'une silhouette de zone de peau non-déformable. Cependant, ses changements d'orientations sont également porteurs de sens et il pourra être nécessaire de les mesurer pour parvenir à une interprétation de plus haut niveau.

3. Les coudes sont, la majorité du temps, situés à l'extérieur du buste et pourront donc être identifiés dans la silhouette du signeur.

4. Les mains sont la plupart du temps situées devant le buste et ne peuvent pour cette raison pas être détectées dans la silhouette du signeur. Il faudra donc utiliser l'information de couleur pour les détecter.

5. Les mains s'occultent fréquemment et occultent également souvent la tête. Pour cette raison, il sera nécessaire de proposer un modèle de posture qui autorise ces contacts dans l'image projetée.

6. Les mains ont une position relative dépendant de main dominante. Cette information peut être utile pour effectuer une désambiguïsation des mains droite et gauche.

6.3 Structure de l'algorithme de suivi

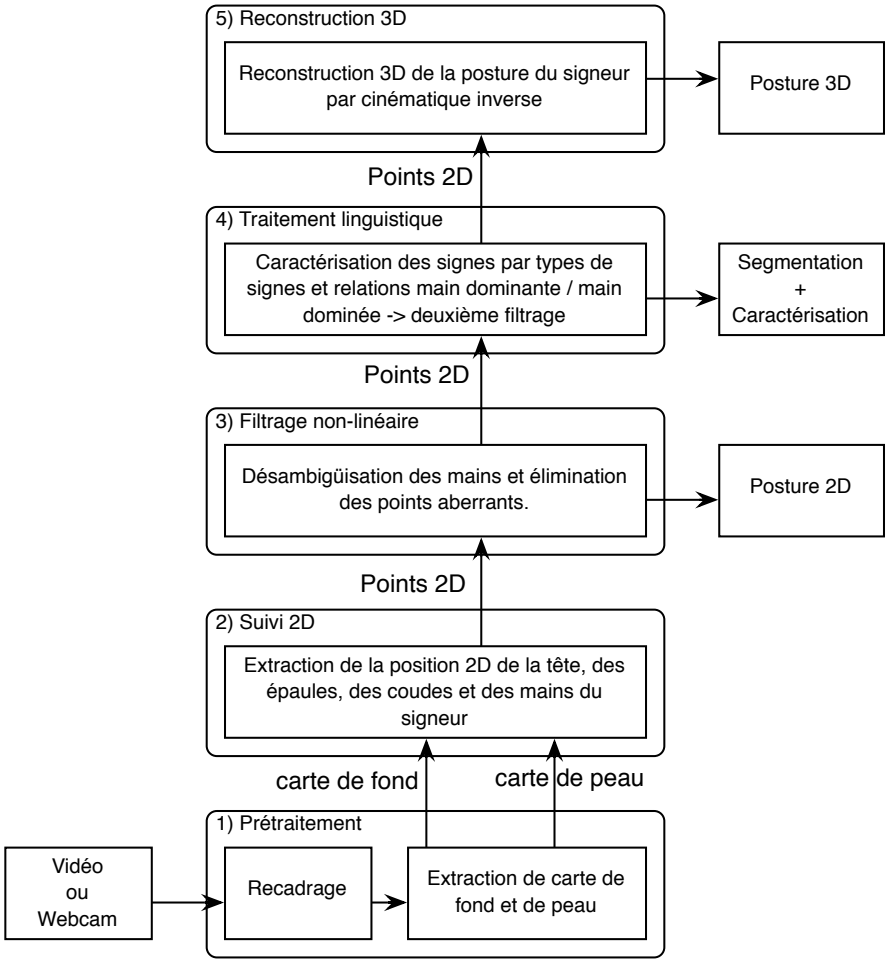


FIGURE 6.17 – Structure générale de l'algorithme de suivi

L'algorithme de suivi que nous avons mis en place se décompose en plusieurs couches de traitement (cf fig. 6.17) que nous allons détailler dans les chapitres qui suivent :

1. La première couche de suivi prend en entrée les images provenant d'un fichier vidéo ou d'un flux vidéo. Après un recadrage éventuel de l'image, elle extrait des cartes de détection de peau et de fond.

2. La seconde couche effectue une détection des différents membres séparément en se basant sur les images de peau et de fond. Elle fournit donc à chaque instant un vecteur de positions 2D des centroïdes de la tête, du buste, des coudes et des mains dans chaque l'image.

3. La troisième couche effectue une désambiguïsation des mains et une rectification des points aberrants.

4. La quatrième couche réalise une caractérisation des signes qui sera utilisée pour réaliser un deuxième filtrage orienté par la phonologie des signes catégorisés.

5. La cinquième couche permet de reconstruire la position 3D du signeur à partir des positions 2D des différents membres.

Nous détaillerons dans ce chapitre le fonctionnement des couches 1, 2, 3 et 5. La couche 4 sera présentée en 8.2.3 car elle fait intervenir des modèles de mouvements de plus haut niveaux.

6.4 Détection de la peau et du fond

Rappelons certaines des hypothèses mentionnées au 4.2.2 portant sur les conditions d'intelligibilité optimale des LS :

- Le fond doit être clair et relativement uni,
- Le signeur doit porter des vêtements unis,
- Le signeur doit être correctement éclairé.

Ces trois contraintes rendent difficile l'utilisation de motifs de texture, mais permettent d'extraire facilement des cartes de détection de peau et de fond puisque les couleurs de la peau, du fond et des habits du signeur sont *a priori* relativement différentes.

6.4.1 Une approche bayesienne

Nous optons pour une approche bayesienne pour estimer l'appartenance d'un pixel de l'image au fond ou à une zone de peau à partir de la couleur d'un pixel. Nous illustrons dans les lignes qui suivent la démarche adoptée pour extraire les pixels de peau. Nous adopterons les notations suivantes :

$\mathbf{color}$: Couleur d'un pixel,

$\mathbf{p}(\mathbf{color})$: Probabilité qu'un pixel de l'image contienne la couleur $color$,

$\mathbf{p}(\mathbf{skin}|\mathbf{color})$: Probabilité qu'un pixel soit dans une zone de peau, sachant sa couleur,

$\mathbf{p}(\mathbf{skin})$: Probabilité qu'un pixel de l'image appartienne à une zone de peau,

$\mathbf{p}(\mathbf{color}|\mathbf{skin})$: Probabilité qu'un pixel ait la couleur $color$ sachant qu'il appartient à une zone de peau.

La règle d'inversion de Bayes nous permet d'écrire :

$$p(skin|color) = p(color|skin).p(skin)/p(color)$$

Dans le cas qui nous intéresse, on fournit à l'algorithme une image de peau constituée d'une arlequine d'images de peau de différentes personnes. L'histogramme de couleur nous permet d'obtenir la probabilité $p(color|skin)$. La morphologie de la personne à suivre a été entrée dans l'algorithme si bien que la probabilité $p(skin)$ peut être calculée comme un rapport entre la surface de l'image et la surface cumulée de la tête et des mains (voire des avants bras si le signeur porte des manches courtes). La probabilité $p(color)$ devrait être calculée à partir de l'histogramme de couleur de l'image courante. Pour des raisons de rapidité d'exécution de notre programme de suivi, nous émettons l'hypothèse que l'histogramme des couleurs des images successives de la vidéo varie relativement peu. Ceci nous permet de calculer $p(color)$ sur une image représentative de la vidéo et d'estimer $p(skin|color)$ avant le début du suivi. Une fois $p(skin|color)$ déterminée, il peut être intéressant d'effectuer un seuillage pour éviter des fausses détections de peau.

6.4.2 Schéma détermination de la couleur de peau et de fond

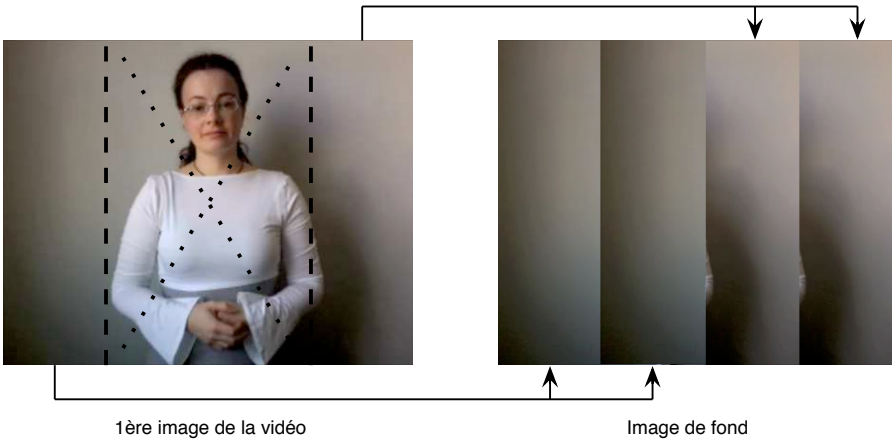


FIGURE 6.18 – Création automatique d'une image de fond à partir de la première image de la vidéo

D'une manière générale, une image de peau standard peut être utilisée dans le cadre de suivi dans des vidéos où le signeur a une teinte de peau proche du modèle de type caucasien et où l'éclairage est suffisant. Dans le cas contraire, il est possible de constituer une image de peau à partir de portions d'images de la vidéo contenant exclusivement la peau du signeur.

Nous suivons une méthode similaire pour la détermination de la probabilité d'appartenance d'un pixel au fond. Il est cependant nécessaire de choisir judicieusement l'image de fond. Dans le cas général,

une image de fond composée des bords droits et gauches de l'image peut être utilisée (fig. 6.18) car les vidéos sont généralement cadrées de manière à ce que le signeur soit sur la première image de la vidéo, au centre du cadre, et en position de repos. Il se pose cependant parfois des problèmes de traitement des ombres portées du signeur sur le fond. Pour simuler la présence d'ombre dans le fond, nous représentons l'image dans l'espace HSV puis nous lui ajoutons un bruit gaussien (majoritairement ajouté sur la composante V la plus affectée par le changement de luminosité). La chaîne utilisée pour l'ajout du bruit est détaillée figure 6.19.
La détermination de la probabilité d'appartenance *a posteriori* d'un pixel à la carte du fond à partir de sa couleur est obtenue comme pour la main en utilisant la règle d'inversion de Bayes.

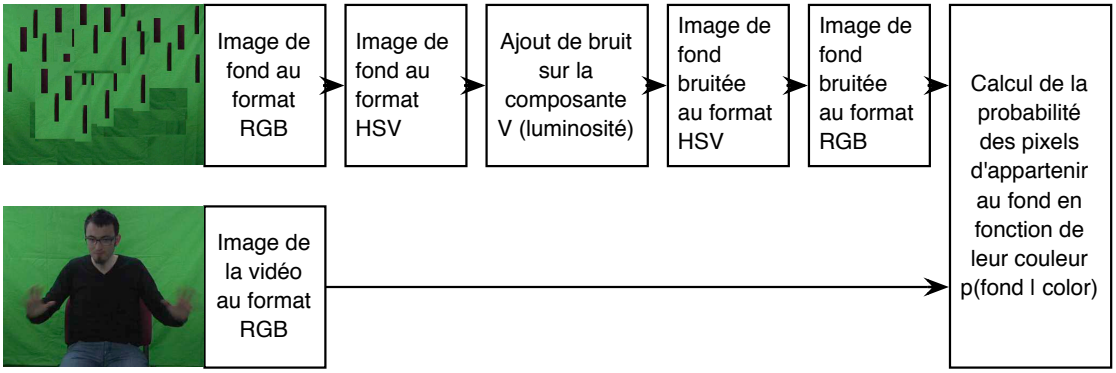


FIGURE 6.19 – Calcul de la probabilité d'appartenance d'un pixel au fond connaissant sa couleur.

6.5 Suivi des différents membres

La méthode de suivi que nous avons mis en œuvre réunit les avantages des filtres particulaires avec partitionnement et des images intégrales.

6.5.1 Retour sur l'état de l'art

Plusieurs remarques peuvent être formulées à propos des différentes méthodes de reconstruction de la posture que nous avons mentionnées dans notre état de l'art :
D'une part, les modèles d'apparence utilisés sont souvent relativement complexes (sphère, cylindres, ellipsoïdes). La morphologie du signeur est souvent bien différente de ces modèles et peu prévisible. Les facteurs engendrant un écart entre le modèle et l'apparence réel sont innombrables (morphologie du signeur, coiffure, forme du col, flexion des membres, longueur des manches ...).
D'autre part, l'espace des postures étant de grande dimension, il est nécessaire de le parcourir de manière judicieuse, de manière à augmenter la vitesse de l'algorithme.

6.5.2 Modèle d'apparence

Partant du constat que la complexité des modèles d'apparence mis en œuvre dans le cas des algorithmes de suivi était souvent superflue, nous utilisons des modèles d'apparence les plus élémentaires possible, ce qui nous permet d'utiliser les images intégrales à l'instar de [MOB06]. Ainsi, chaque partie du corps est modélisée par une forme géométrique aux bords verticaux et horizontaux. L'image 6.20 récapitule l'apparence des différents membres. Nous détaillons ensuite, dans la section 6.5.2.1, le fonctionnement des différents filtres.

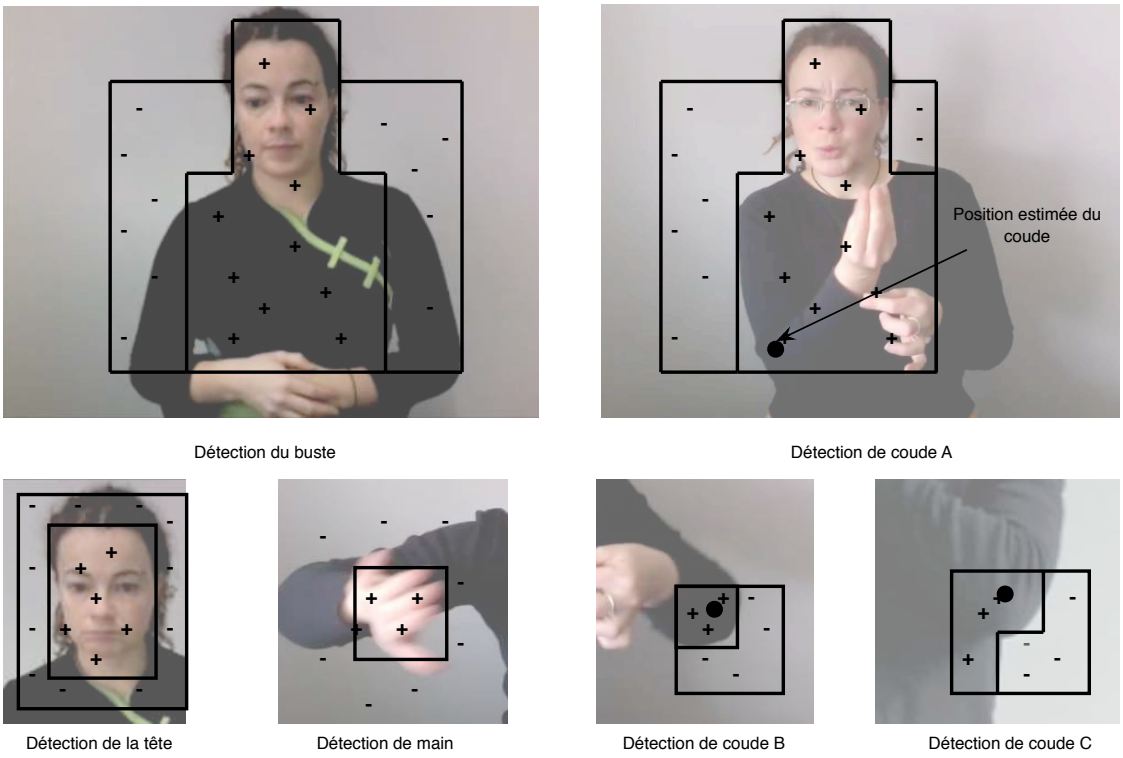


FIGURE 6.20 – Filtres utilisés pour la détection et la localisation des parties du corps.

6.5.2.1 Modèle de la tête et du buste

Prenons l'exemple de la tête du signeur pour détailler le fonctionnement des filtres. Nous modélisons son apparence par celle de deux rectangles emboîtés (fig. 6.21). Le rectangle intérieur contient une zone de peau. L'espace entre le rectangle intérieur et le rectangle extérieur n'en contient pas. On note $p(x, y)$ la probabilité d'appartenance d'un pixel de coordonnées (x, y) à une zone de peau.

La réponse R_f du filtre appliqué au rectangle ABCD de l'image (fig. 6.21) est la différence de deux

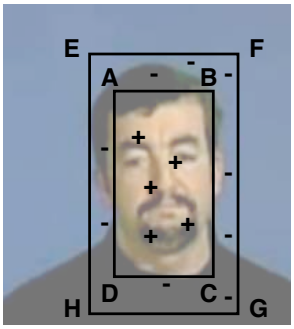


FIGURE 6.21 – Filtre utilisé pour le suivi de la tête du signeur.

intégrales[3] :

$$R_f = \int_{(x,y)\in ABCD} p(x,y).dx.dy - \int_{(x,y)\in EFGH-ABCD} p(x,y).dx.dy$$

$$R_f = 2.\int_{(x,y)\in ABCD} p(x,y).dx.dy - \int_{(x,y)\in EFGH} p(x,y).dx.dy$$

Ces intégrales peuvent aisément être calculées à l'aide d'images intégrales (cf. annexe II). Nous notons $I(P)$ la valeur de l'image intégrale de la carte de détection de peau au point P.

$$R_f = 2.(I(A) - I(B) + I(C) - I(D)) - (I(E) - I(F) + I(G) - I(H))$$

Le résultat du filtrage est donc extrêmement économique en temps de calcul.

Nous suivons une approche similaire pour la détection du buste dont le modèle est visible fig. 6.20.

6.5.2.2 Détection des coudes

Comme nous l'avons vu en 6.2.4, les coudes ont une position quasi-stationnaire à l'extérieur du buste, ce qui leur donne une apparence peu variable. Les premiers tests que nous avons fait en utilisant un seul filtre de détection provoquaient des décrochages, lorsque la silhouette du coude se trouvait totalement ou partiellement masquée par le buste du signeur. Ceci nous obligeait ensuite à utiliser une méthode de filtrage non-linéaire décrite en 6.7 pour supprimer les points aberrants de suivi du coude. Nous avons donc réalisé plusieurs filtres adaptés à différentes situations (cf. figure 6.20).
Les filtres ne sont pas appliqués en cascade comme c'est le cas dans les filtres intégraux utilisés pour la détection de visage par [VJ01] où les résultats des différents filtres sont fusionnés par moyenne pondérée.

[3]Les résultats du filtrage sont normalisés par la surface du filtre $S_{ABCD} = S_{EFGH}/2$ pour que la réponse du filtre soit comprise entre –1 et 1.

Dans notre cas en revanche, les différentes apparences du coude sont mutuellement exclusives et nous utilisons par conséquence une fusion de type MAX.
La position du coude est cependant légèrement erronée lorsque le coude passe devant le buste du signeur car il est alors toujours localisé par les filtres sur la bordure de la silhouette du buste. La position du coude sera alors corrigée lors du filtrage bas niveau (voir section 6.7).

6.5.2.3 Détection des mains

Le suivi des mains du signeur est un problème particulier qui peut difficilement être dissocié de celui du suivi de la tête. Nous nous plaçons dans un premier temps, dans le cas d'un signeur à manche longue. Le modèle d'apparence retenu est de représenter chaque membre (tête et mains) par un rectangle. Le résultat du filtre est donné par la corrélation entre le modèle et la carte de peau dans l'ensemble de l'image.
Dans nos calculs, nous appellerons $var(Mod)$ la variance du modèle de peau, $var(Cdp)$ la variance

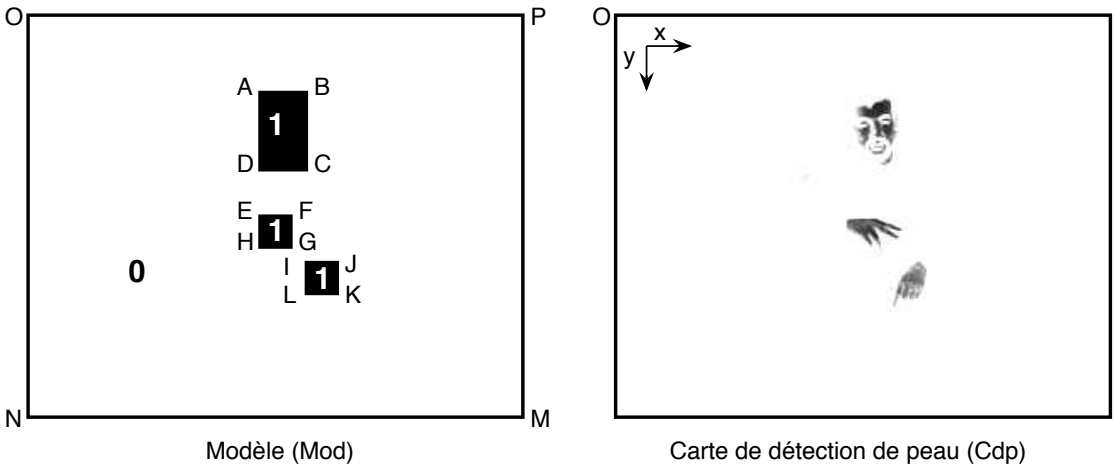


FIGURE 6.22 – Filtre utilisé pour le suivi de la tête du signeur.

de la carte de détection de peau et $covar(Mod, Cdp)$ la covariance entre le modèle et la carte de détection de peau. La surface de l'image MNOP sera notée S. Pour alléger les notations, nous noterons Δ le domaine recouvert par l'union des rectangles ABCD, EFGH et IJKL et S_Δ sa surface. La corrélation $correl$ entre le modèle et la carte de détection de peau est déterminée de la manière suivante :

$$correl = covar(Mod, Cdp)/\sqrt{var(Mod).var(Cdp)}$$

Les corrélations peuvent être déterminées par les intégrales suivantes dans lesquelles $q(x, y)$ et $p(x, y)$ représentent respectivement les valeurs des pixels des images Mod et CdP au point de coor-

données (x,y) :

$$covar(Mod, Cdp) = \frac{1}{S} \int_{(x,y)\in MNOP} (q(x,y) - \mu_q).(p(x,y) - \mu_p).dx.dy$$

$$covar(Mod, Cdp) = \frac{1}{S} \int_{(x,y)\in MNOP} q(x,y).p(x,y).dx.dy - \mu_q.\mu_p$$

$$covar(Mod, Cdp) = \frac{1}{S} \int_{(x,y)\in \Delta} p(x,y).dx.dy - \frac{S_\Delta}{S}.\mu_p$$

$$covar(Mod, Cdp) = \frac{1}{S} \int_{(x,y)\in \Delta} p(x,y).dx.dy - \frac{S_\Delta}{S^2} \int_{(x,y)\in MNOP} p(x,y).dx.dy$$

$$var(Mod) = S_\Delta.(S - S_\Delta)/S^2$$

$$var(Cdp) = Cste$$

Nous ne prenons pas la peine de déterminer précisément la variance de la carte de détection de peau $var(Cdp)$, car elle ne sera pas affectée par la variation du modèle. Il est naturellement possible d'utiliser de nouveau les images intégrales afin d'optimiser le temps de calcul :

$$covar(Mod, Cdp) = \frac{1}{S}(\ I(A) - I(B) + I(C) - I(D) + I(E) - I(F) + I(G) - I(H)$$

$$+I(I) - I(J) + I(K) - I(L)\ ) - \frac{S_\Delta}{S^2} I(M)$$

Notons que les rectangles correspondant aux mains et à la tête peuvent se recouvrir partiellement ou totalement. Il est aisé de tenir compte des intersections dans les calculs d'intégrales car cela ne nécessite que d'effectuer des intersections de rectangles. Le modèle est également assez flexible pour que les mains puissent sortir de l'image. Ceci ne change en rien la formule générale de calcul de corrélation.

La surface de peau théorique générée par notre modèle est variable suivant que les zones correspondant à la tête et aux deux mains se recouvrent ou non. Nous avons ajouté un facteur α à la formule de corrélation exposée pour sur-noter les hypothèses de posture où la surface de recouvrement entre les différents membres est minimale. Cela permet d'améliorer légèrement la robustesse du suivi en évitant les situations où les mains sont injustement détectées au même endroit que la tête. La formule de calcul de corrélation entre le modèle et la carte de détection de peau devient alors :

$$correl = covar(Mod, Cdp)/\sqrt{K_v.S_\Delta.(S - \alpha.S_\Delta)/S^2}$$

6.5.2.4 Détection des avant-bras

Nous avons ensuite généralisé notre méthode à des signeurs dont l'avant-bras est visible. Il n'est alors plus possible de modéliser l'ensemble de l'avant-bras par un unique rectangle. Nous choisissons donc d'approximer la forme de l'avant-bras par un ensemble de rectangles de manière à être plus précis, tout en conservant la rapidité de l'algorithme (fig. 6.23). La formule utilisée pour le calcul de corrélation entre le modèle et la carte de peau est en tout point identique à la formule mentionnée en 6.5.2.1.

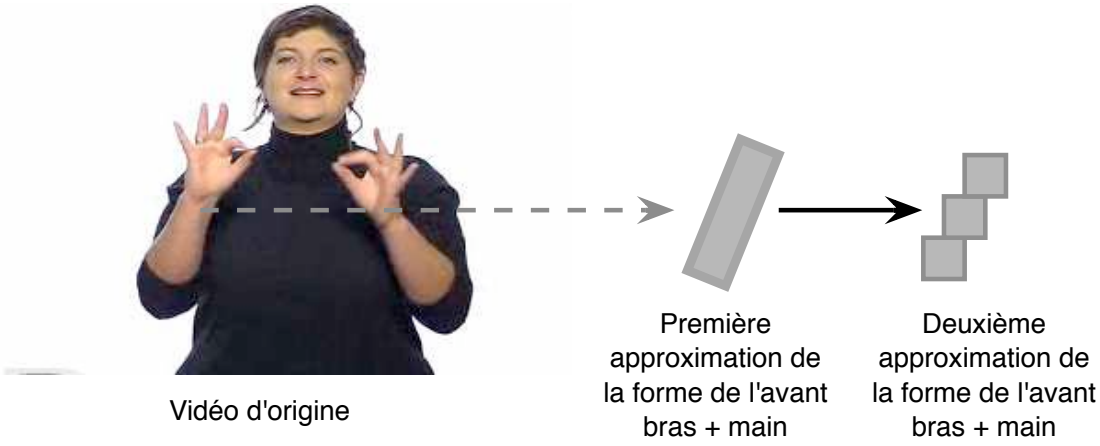


FIGURE 6.23 – Approximation de la forme des avant bras compatible avec l'utilisation d'images intégrales.

6.5.3 Mise en place d'un filtrage particulaire partitionné

Le suivi des différents membres est réalisé grâce à un filtre particulaire. Nous suggérons au lecteur désirant en savoir plus sur les différents types de filtres particulaires de consulter l'annexe I.
Nous cherchons à déterminer la position 2D de la tête, du buste, des coudes et des mains du signeur dans la vidéo. Cela représente en tout 12 coordonnées. Le temps de parcours de l'espace de paramètre par un filtre particulaire augmente exponentiellement avec la dimension de l'espace à parcourir. Nous avons donc opté pour un partitionnement de l'espace des paramètres, permettant de parcourir successivement les différentes dimensions.
Nous décrivons dans les sections qui suivent, comment ce partitionnement du filtre particulaire a été réalisé, puis nous détaillons le modèle de dynamique des mouvements utilisé ainsi que le la méthode de détermination des poids des différentes particules.

6.5.3.1 Le partitionnement

Le parcours optimal de l'espace des postures est obtenu en mettant en œuvre un filtre particulaire partitionné à 5 partitions (cf. annexe I). Cela permet d'intégrer les contraintes cinématiques directement

dans l'algorithme de suivi. Chaque filtre particulaire est basé sur un ou plusieurs filtres comme ceux décrits au §6.5.2.1 qui sont appliqués sur une image intégrale. L'image intégrale est, soit calculée sur l'image de détection de peau, soit sur l'image de détection du fond. Chaque filtre particulaire est également caractérisé par une dimension (nombre de coordonnées des membres à suivre) et des contraintes par rapport aux positions des membres qui ont déjà été suivis. Nous avons implémenté les contraintes suivantes :

- La tête doit être située au moins partiellement dans la vidéo,
- Le buste se trouve en dessous de la tête,
- La distance entre les épaules et les coudes correspondants est inférieure à la longueur du bras.

Ces contraintes permettent à la fois de réduire la taille de l'espace de recherche et de supprimer une partie des postures aberrantes. Le tableau 6.24 résume les caractéristiques des différents filtres.

Membres suivis	Entrée	filtre utilisé	Dim	Sortie	Contrainte
Tête	Peau	§6.5.2.1	2	x_{tete}, y_{tete}	Tête dans l'image
Buste	Fond	§6.5.2.1	2	x_{buste}, y_{buste}	buste sous la tête
Coude droite	Fond	§6.5.2.2	2	x_{coudeD}, y_{coudeD}	D_{max} épaule-coude
Coude gauche	Fond	§6.5.2.2	2	x_{coudeG}, y_{coudeG}	D_{max} épaule-coude
Mains	Peau	§6.5.2.3	4	x_{mainD}, y_{mainD} x_{mainG}, y_{mainG}	aucune

FIGURE 6.24 – Résumé des différents filtres particulaires résultant du partitionnement

6.5.3.2 Modèle de dynamique

En raison du grand nombre de points de rebroussement dans les trajectoires des mains, nous avons choisi d'utiliser un modèle de déplacement de type marche aléatoire. Le déplacement des particules peut donc être modélisé par un bruit gaussien ajouté à chacune de leurs coordonnées. L'amplitude du bruit gaussien est une fonction de plusieurs paramètres :

- Plus le cadrage de la vidéo est large, moins le déplacement apparent des membres dans la vidéo est important.
- La dynamique dépend du membre à suivre. Ainsi par exemple, elle est plus importante pour les mains que pour le buste.
- La variation de la position des membres est aussi dépendante du signeur et du type de production. Il est donc également nécessaire de paramétrer la dynamique des membres en fonction du *style* du signeur.

6.5.3.3 Détermination du poids des particules

Nous utilisons une démarche proche de [Fon08] pour estimer le poids des particules w à partir des mesures de corrélation $correl$ dont le mode de calcul est exposé en 6.5.2.3.

$$w = e^{correl*\sigma}$$

Cette formule fait apparaître un paramètre σ qui indique la sélectivité du filtre particulaire. Ce paramètre a une certaine influence sur la précision des résultats, mais est fixé expérimentalement, faute de méthode adéquate pour le déterminer automatiquement. Nous préciserons l'influence du paramètre σ dans l'évaluation du §6.10.3.

6.6 Désambiguïsation des mains

Les statistiques mentionnées au §6.2.5 font apparaître de fréquentes inversions des deux mains. Dans certaines productions, nous avons relevé [LAD09a] des croisements dans près de 10% des images de la vidéo. Ce problème peut être résolu au niveau des filtres particulaires si le signeur porte des manches courtes car l'avant-bras permet de relier la main au coude correspondant. Cependant, les deux mains sont indistinguables (d'après notre modèle) lorsque le signeur porte des manches longues. D'autres critères doivent alors être pris en compte pour effectuer la désambiguïsation.

Nous savons d'après les statistiques du §6.2.5 que la position relative des mains est un critère de désambiguïsation utilisable. Nous savons par ailleurs que la distance entre le coude et la main correspondante, projetée sur le plan de l'image, admet une borne supérieure. Enfin, la continuité du mouvement se traduit par un déplacement limité de chaque main entre deux images consécutives de la vidéo. Nous associons une mesure de vraisemblance à chacun des critères que nous venons de décrire. Les sections qui suivent décrivent comment sont déterminées ces mesure de vraisemblance[4].

6.6.1 Utilisation de la position relative des mains

Les statistiques qui suivent ont été établies à partir d'une traduction de brèves d'actualité en LSF de 30s fournie de Websourd et vont dans le même sens que les mesures que nous avons obtenues à partir des données de capture de mouvement. Nous nommerons $H1$ et $H2$ les deux mains et nous noterons $x_{H1}(t), y_{H1}(t)$ et $x_{H2}(t), y_{H2}(t)$ leurs coordonnées à l'instant t. Nous déterminons par une méthode bayésienne les deux mesures suivantes :

[4]Nous préférons parler de mesures de vraisemblance plutôt que de probabilités dans notre cas, car nous ne maîtrisons pas la dépendance des différents critères utilisés pour la désambiguïsation.

– P_1 représente la mesure de vraisemblance que la main $H1$ soit la main droite en connaissant la différence d'abscisse $x_{H2}(t) - x_{H1}(t)$.

– P_2 représente la mesure de vraisemblance que la main $H1$ soit la main droite en connaissant la différence d'ordonnée $y_{H2}(t) - y_{H1}(t)$.

Les résultats sont présentés dans les figures 6.25 et 6.26. Nous modélisons chacune de ces mesures de vraisemblance par une sigmoïde. Ainsi, les mesures de vraisemblance P_1 et P_2 sont déterminées grâce aux formules suivantes :

$$P_1(t) = 1/[1 + e^{\lambda_1.(x_{H1}(t) - x_{H2}(t))}]$$

$$P_2(t) = 1/[1 + e^{\lambda_2.(y_{H1}(t) - y_{H2}(t))}]$$

Les paramètres λ_1 et λ_2 sont déterminés dynamiquement lors de l'effectuation du suivi (cf. §6.9).

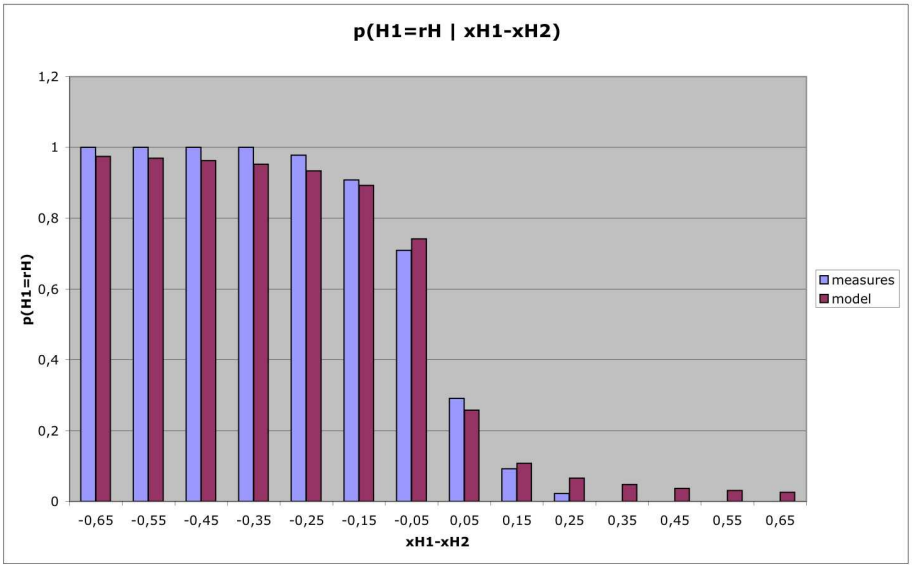


FIGURE 6.25 – Modélisation de P_1

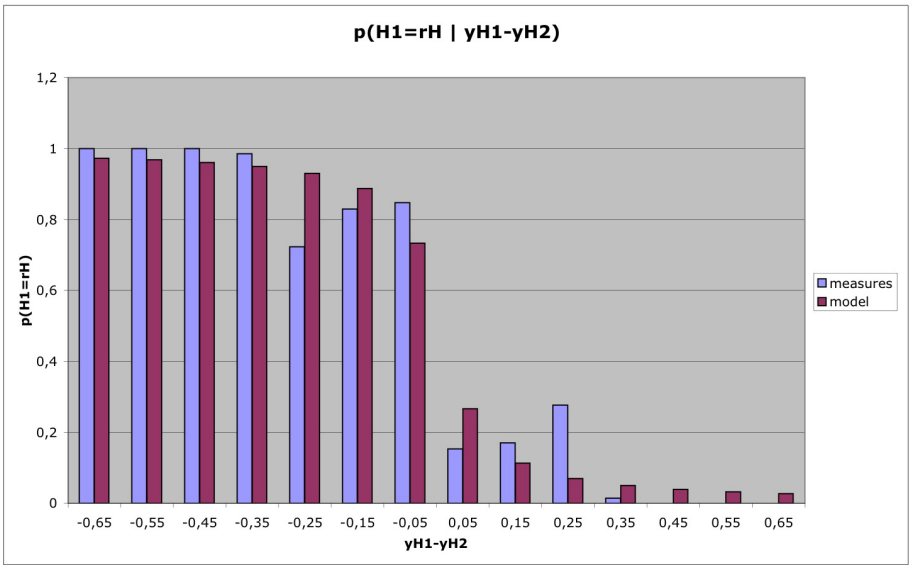


FIGURE 6.26 – Modélisation de P_2

6.6.2 Utilisation de la distance main coude

Nous savons que la position de la main est contrainte par la chaîne cinématique qui la relie au coude. Nous utilisons donc également la distance main coude comme critère de désambiguïsation. Nous appelons $P_3d(t)$ la mesure de vraisemblance que la main $H1$ soit la main droite à l'instant t en connaissant la distance main droite – coude droit. Nous déterminons cette mesure de vraisemblance par une méthode bayésienne comme précédemment. Les résultats sont représentés sur la figure 6.27 où $dist(H1, re)$ désigne la distance entre la main droite et le coude droit.

6.6.3 Utilisation de la continuité du mouvement

Le dernier critère que nous prenons en compte pour la désambiguïsation des mains est le déplacement de chaque main entre deux images successives. La vraisemblance des différents déplacements découle directement du modèle de dynamique que nous utilisons. Soit $\Delta_d(t, t+1)$ le déplacement de la main droite entre deux images successives, la vraisemblance de l'observation sera calculée de la manière suivante :

$$P_4d(t, t+1) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\Delta_d^2(t,t+1)}{2\sigma^2}}$$

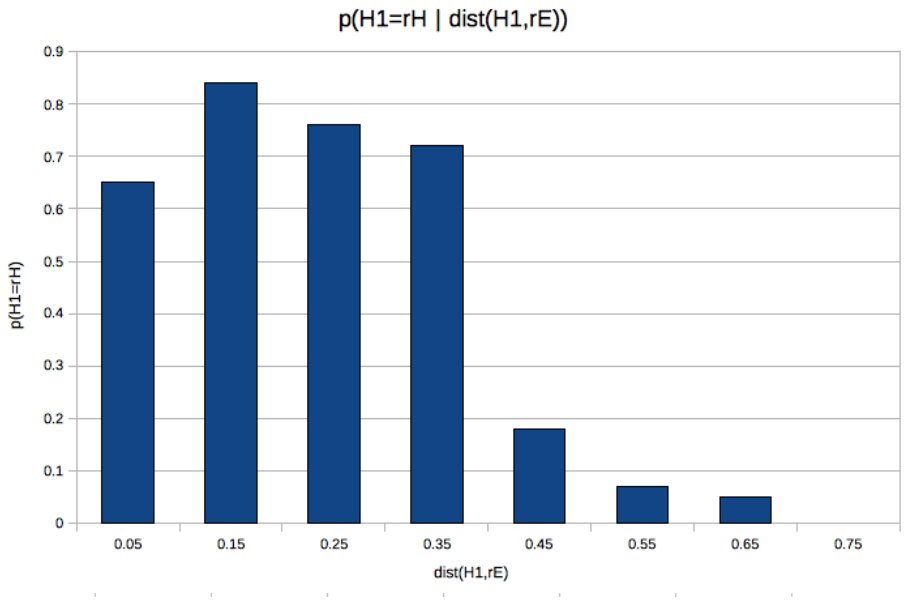


FIGURE 6.27 – Modélisation de P_3d

Notons que, contrairement aux mesures de vraisemblances précédentes qui peuvent être calculées à un instant t, les mesures de vraisemblances relatives au déplacement des mains P_4d et P_4g sont calculées entre deux images successives.

6.6.4 Fusion des mesures de probabilité

Il est désormais nécessaire de fusionner les différentes mesures de vraisemblances. Nous les fusionnons en adoptant une démarche de Fusion Bayesienne Naïve. La mesure de vraisemblance V d'un appariement peut être calculée sur l'ensemble des N images de la vidéo en pondérant les mesures de vraisemblances statiques $P_1(t)$, $P_2(t)$, $P_3d(t)$, $P_4d(t)$, et les mesures de vraisemblance dynamiques $P_4d(t,t+1)$ et $P_4g(t,t+1)$ grâce au facteur α qui est fixé empiriquement ($\alpha = 1$ dans notre implémentation).

$$B(t) = P_1(t).P_2(t).P_3d(t).P_4d(t)$$

$$C(t,t+1) = P_4d(t,t+1).P_4g(t,t+1)$$

$$V = \prod_{t \in N} B(t)^{\alpha}.C(t,t+1)$$

L'appariement correct est celui qui maximise V. La maximisation peut être obtenue rapidement en utilisant des techniques de programmation dynamique. Dans la version temps-réel de l'algorithme, les mesures entre la première image de la vidéo et l'image courante sont utilisées. Dans sa version hors ligne, l'ensemble des mesures du suivi des coudes et des mains dans la vidéo est utilisé pour optimiser la désambiguïsation.

6.7 Correction de la position des coudes

Nous avons mentionné en 6.5.2.2, le fait que la position des coudes détectée par les filtres particulaires est parfois légèrement erronée lorsque le coude se trouve devant le buste du signeur. Heureusement, cette situation correspond souvent à une posture où la main doit atteindre une cible éloignée du côté opposé. Nous utilisons donc la contrainte cinématique liée à la longueur de l'avant-bras pour trouver la posture satisfaisant l'ensemble des contraintes cinématiques, la plus proche de la posture déterminée par le filtrage. Ceci permet en général de réduire l'erreur d'estimation de la position des coudes.

6.8 Correction de l'estimation de la profondeur des mains du signeur

Dans un premier temps, nous décidons de déterminer la position 3D des mains par rapport aux épaules en utilisant uniquement les positions des mains, des coudes et des épaules mesurées sur l'image. Le corps humain ayant un nombre important de degrés de mobilité, nous simplifions la chaîne cinématique modélisant le corps pour mener à bien les calculs de détermination de la profondeur.
Nous émettons l'hypothèse que le buste, les épaules et la tête du signeur sont tous dans le même plan Π. Dans les calculs qui suivent, la profondeur Z sera orthogonale à ce plan[5]. Les bras et avant bras sont modélisés comme des solides indéformables ; les coudes et les épaules sont considérés comme des liaisons pivot sans limitation d'angle. Notre modèle ne gère pas les gestions de collision.
Comme nous le mentionnions dans le §4.2.8, le problème d'estimation des différences de profondeur entre la main et l'épaule en connaissant la position 2D de la main, de l'épaule et du coude et en connaissant la longueur des bras admet quatre solutions dans le cas général. Les statistiques que nous avons réalisées dans le §6.2.5 sur les positions relatives des coudes des épaules et des mains montrent que le cas le plus fréquent est $Z_s < Z_e < Z_h$ où Z_s, Z_e et Z_h représentent respectivement les profondeurs relatives des mains (indice $_h$), des coudes (indice $_e$) et des épaules (indice $_s$).
La reconstruction de la posture en utilisant uniquement la cinématique inverse fournit des résultats extrêmement bruités. Pour cette raison, nous choisissons d'effectuer plusieurs types de corrections décrites en détail dans [LAD09c] :

1. Un lissage à l'aide d'un Filtre de Kalman Etendu,

2. Une amélioration de l'estimation de la profondeur en utilisant la corrélation des profondeurs des deux mains.

[5]Nous négligeons donc les différences de profondeur entre les deux épaules causées par les rotations du buste qu'on peut observer dans certaines structures de transfert.

Les équations menant à la détermination de la profondeur à partir de positions 2D des articulateurs sont non-linéaires. Pour cette raison, nous utilisons un lisseur de Kalman étendu pour obtenir à chaque pas de temps, une estimation optimale de la profondeur de chaque main et de la covariance de chacune de ces estimations.

Il est fréquent que la main dominée du signeur soit en position de repos. Dans ce cas, la main est souvent proche du corps, ce qui conduit à une différence de profondeur faible entre le coude et la main du signeur. L'estimation de la profondeur de la main est alors entachée d'une énorme incertitude. Par ailleurs, nous savons que les mouvements des deux mains en profondeur ne sont pas indépendants. Nous obtenons une corrélation de l'ordre de 0,24 entre les profondeurs des deux mains. Plus qualitativement, on observe que la main dominée reproduit souvent les mouvements de la main dominante, même dans les signes où une seule main est censée être en mouvement d'après les modèles linguistiques. Nous décidons d'utiliser la corrélation entre les profondeurs des mains droite et gauche de manière à améliorer les estimations de la profondeur de la main dominée. A l'issue de l'utilisation du lisseur de Kalman, nous obtenons :
La profondeur de la main dominante (indice $_s$) : $Z_s^a \sim N(\mu_{Z_s^a}, \sigma_{Z_s^a})$
La profondeur de la main dominée (indice $_w$) : $Z_w^a \sim N(\mu_{Z_w^a}, \sigma_{Z_w^a})$
Les exposants a et b désignent respectivement les estimations de profondeur avant et après l'amélioration prenant en compte la corrélation des profondeurs. L'estimation corrigée de la profondeur de la main gauche est notée $Z_w^b \sim N(\mu_{Z_w^b}, \sigma_{Z_w^b})$. Elle est obtenue de la manière suivante :

$$\mu_{Z_w^b} = \mu_{Z_w^a} + \frac{(\sigma_{Z_w^a})^2}{(\sigma_{Z_w^a})^2 + (\sigma_{Z_s^a})^2.\beta}.(\mu_{Z_s^a} - \mu_{Z_w^a})$$

$$(\sigma_{Z_w^b})^2 = (\sigma_{Z_w^a})^2 - \frac{(\sigma_{Z_w^a})^2}{(\sigma_{Z_w^a})^2 + (\sigma_{Z_s^a})^2.\beta}.(\sigma_{Z_w^a})^2$$

Le paramètre β est estimé en fonction de la dépendance entre les profondeurs des deux mains. Si ces deux mesures sont totalement indépendantes, cette valeur est $+\infty$. Si au contraire, les deux mains ont constamment la même profondeur, cette valeur est 1. Dans notre expérience, la valeur $\beta = 2.5$ a mené aux meilleurs résultats.

La formule que nous utilisons pour l'amélioration de l'estimation de la profondeur de la main dominée est inspirée par le filtre de Kalman. Il serait possible de l'optimiser en utilisant simultanément les mesures des profondeurs de la main dominante et de la main dominée à instants t et leurs estimations à l'instant $t-1$ comme quatre mesures d'un même système dont on cherche connaître l'état à l'instant t.

6.9 Vers un paramétrage automatique

L'ensemble des techniques de filtrage que nous venons d'énumérer, fait appel à un certain nombre de paramètres liés à l'apparence du signeur, à la distribution de la position relative des mains, des distances main-coude et à la dynamique du mouvement.
Il est naturellement possible de fixer les paramètres avant le suivi, mais ceci nécessite de connaître les caractéristiques du signeur à suivre. Un re-paramétrage systématique de l'algorithme avant chaque nouveau suivi est alors nécessaire. Ceci implique parfois une approche par essai-erreur et une intervention constante d'un opérateur humain.
Nous avons donc essayé d'optimiser cette phase d'apprentissage des paramètres. Pour l'instant, les paramètres de désambiguïsation des mains sont les seuls à être appris automatiquement. Le mécanisme que nous avons utilisé est une mise à jour des paramètres des distributions au cours du suivi. Par contre, l'initialisation de la morphologie du signeur est encore manuelle.

6.10 Évaluation

Nous évaluerons plusieurs caractéristiques de notre algorithme de suivi à partir d'une vidéo de deux minutes dans laquelle un signeur raconte un fait marquant d'actualité. Les vérités terrains avec lesquelles sont comparées notre suivi ont été saisies manuellement par F. Gianni [Gia08] en cliquant directement les positions des différents membres sur chaque image de la vidéo.
Nous évaluons dans cette partie, la précision du suivi des mains du signeur en fonction du paramétrage des filtres particulaires et la précision de la reconstruction 3D de la posture du signeur. Par ailleurs, nous évaluerons les performances de notre algorithme en terme de temps de calcul.

6.10.1 Rapidité du suivi

La rapidité du suivi pour traiter une vidéo de 2 minutes est évaluée en fonction du nombre de particules. Comme le montre le graphique 6.28, ce temps augmente linéairement avec le nombre de particules. On note toutefois un temps minimal de traitement de la vidéo qui est probablement dû aux calculs nécessaires pour décompresser la vidéo.

6.10.2 Efficacité de la désambiguïsation des mains

Nous avons fait varier le nombre de paramètres pris en compte pour effectuer la désambiguïsation des mains et noté, pour chaque combinaison, le nombre d'erreurs de suivi causées par les inversions des mains. Nous utilisons la variante temps réel de l'algorithme de désambiguïsation. Les résultats

Nombre de particules par filtre	temps de traitement (en secondes)
100	50
500	50
1000	50
2000	79
5000	166
10000	303

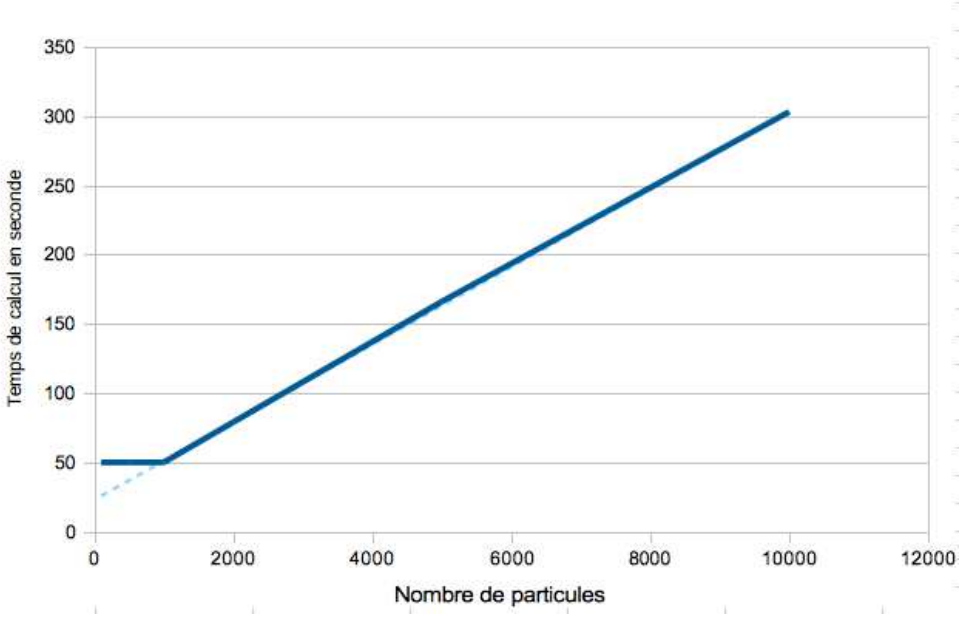


FIGURE 6.28 – Temps de calcul en fonction du nombre de particules

sont exposés dans le tableau 6.29.

Les résultats font ressortir que les différents paramètres utilisés améliorent bien la désambiguïsation des mains du signeur. Toutefois, il semble également que les critères d'abscisse relative des mains et de distance main-coude sont parfois redondants.
Des performances de désambiguïsation encore meilleures peuvent être obtenues si on utilise la version hors ligne de l'algorithme de désambiguïsation des mains. Nous avons atteint seulement 2% d'erreurs dans l'étude [LAD09a] portant sur une vidéo de 750 images.

6.10.3 Robustesse du suivi

La robustesse de l'algorithme est évaluée par rapport au nombre de décrochages de l'algorithme de suivi. Nous désignons par décrochages les instants ou la position estimée d'une main est éloignée de la position réelle de la main de plus de 67 pixels, ce qui correspond à environ deux fois la largeur de la main. Ce critère est aussi utilisé par [Gia08]. Par contre, à la différence de cette dernière étude, nous comptons les erreurs des appariements dans les erreurs de suivi. Les différentes mesures sont ef-

Abscisse	Ordonnée	Distance main-coude	Continuité	Nombre d'erreurs
		X		119
		X	X	113
	X	X		133
	X	X	X	72
X				172
X			X	124
X		X		137
X		X	X	115
X	X			135
X	X		X	100
X	X	X		109
X	X	X	X	68

FIGURE 6.29 – Nombre d'erreurs d'appariements en fonction des paramètres utilisés pour la désambuïsation des mains

fectuées sur les images 300 à 2800 de la vidéo test de près de 3000 images. Les 300 premières images sont nécessaires pour que le filtre de désambiguïsation s'initialise correctement. Les 200 dernières images ne contiennent que la tête du signeur, mais le signeur ne transmettait plus d'informations. Nous faisons varier à la fois σ (la sensibilité des filtres) et le N_{part} (le nombre de particules). Les graphes 6.32 et 6.33 représentent le nombre de décrochages du filtre particulaire sur la main droite en fonction de σ et du nombre de particules N_{part}. Notons que chaque mesure de robustesse n'a été effectuée qu'une seule fois et que les résultats fournis par le filtre particulaire sont susceptibles de varier légèrement d'une fois sur l'autre, du fait de l'utilisation de nombres aléatoires.

Nombre de particules	erreur moyenne (pixel)	part de points aberrants
8	53,84	29,8%
15	38,98	19,8%
25	23,02	7,3%
50	15,48	1,92%
100	13,42	1,12%
250	12,24	0,32%
500	11,80	0,24%
1000	11,63	0,16%
4000	11,69	0,20%

FIGURE 6.30 – Influence du nombre de particules sur l'erreur de suivi

Valeur de σ	erreur moyenne (pixel)	part de points aberrants
7	36,11	13,16%
11	13,9	1,08%
23	**10,7**	**0,08%**
32	11,25	0,08%
38	11,31	0,16%
46	11,63	0,16%
69	12,6	0,16%
91	13,14	0,16%
115	13,24	0,36%
161	14,32	1,32%

FIGURE 6.31 – Influence de σ sur l'erreur de suivi

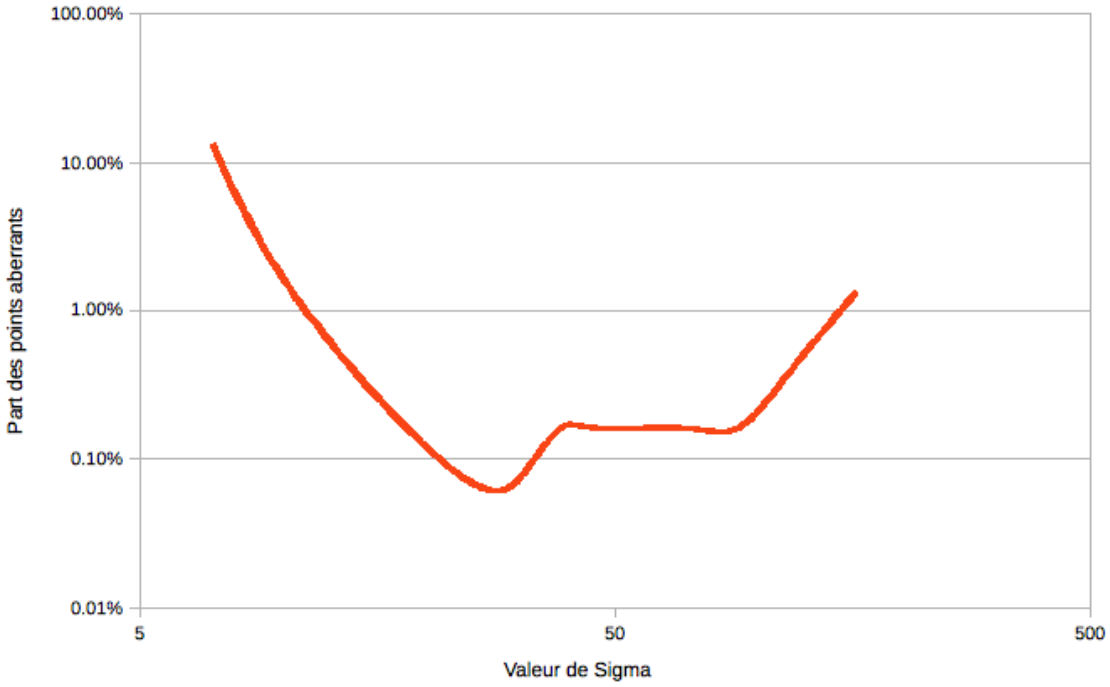


FIGURE 6.32 – Influencede σ sur la part de points aberrants

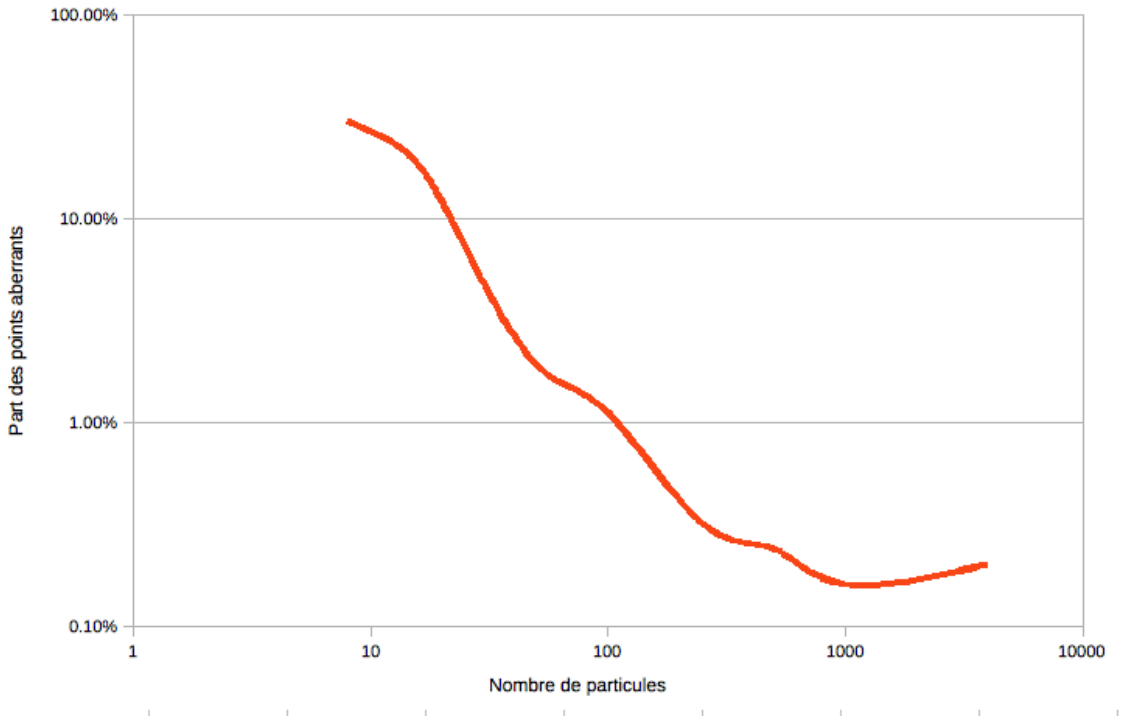


FIGURE 6.33 – Influence du nombres de particules sur la part de points aberrants

De bonnes performances sont atteintes, même en utilisant un faible nombre de particules. On peut estimer que l'utilisation de 1000 particules permet d'obtenir un nombre de décrochages proche du minimum. Le paramètre σ a relativement peu d'influence sur la robustesse.
D'un point de vue qualitatif, les différents essais ont démontré que les décrochages sont en général de courte durée, et que les filtres particulaires rattrapent rapidement leur cible après un décrochage. Les cas d'occultations partielles ne causent que peu de décrochages. Par contre, les cas d'occultations totales d'une main par la tête engendrent parfois des erreurs de suivi. La sortie des mains du cadre de la vidéo est théoriquement supportée par notre modèle de suivi. L'évaluation a en fait montré que les sorties de cadre ne sont gérées de manière satisfaisante que pour une courte durée. Ceci est la raison qui nous a poussé à supprimer les 200 dernières images de notre évaluation.

6.10.4 Précision du suivi

Nous utilisons la même vidéo pour évaluer la précision du suivi. Comme dans l'évaluation de la robustesse, nous ne prenons en compte que les images de 300 à 2800 et notre évaluation porte sur la précision du suivi de la main droite[6].
Le graphe 6.35 présente la répartition des erreurs de suivi évaluées en pixel pour un filtre de $N_{part} =$

[6]Nous avons toutefois pris le soin de vérifier que la répartition des erreurs de suivi étaient la même pour les mains droite et gauche.

1000 particules. A titre indicatif, la largeur de la main est de 34 pixels. La répartition des erreurs est type log-normale. L'erreur moyenne est de 11,3 pixel soit environ un tiers de la paume de la main. Les graphes 6.34 et 6.36 présentent respectivement la précision du suivi en fonction de N_{part} et de σ.

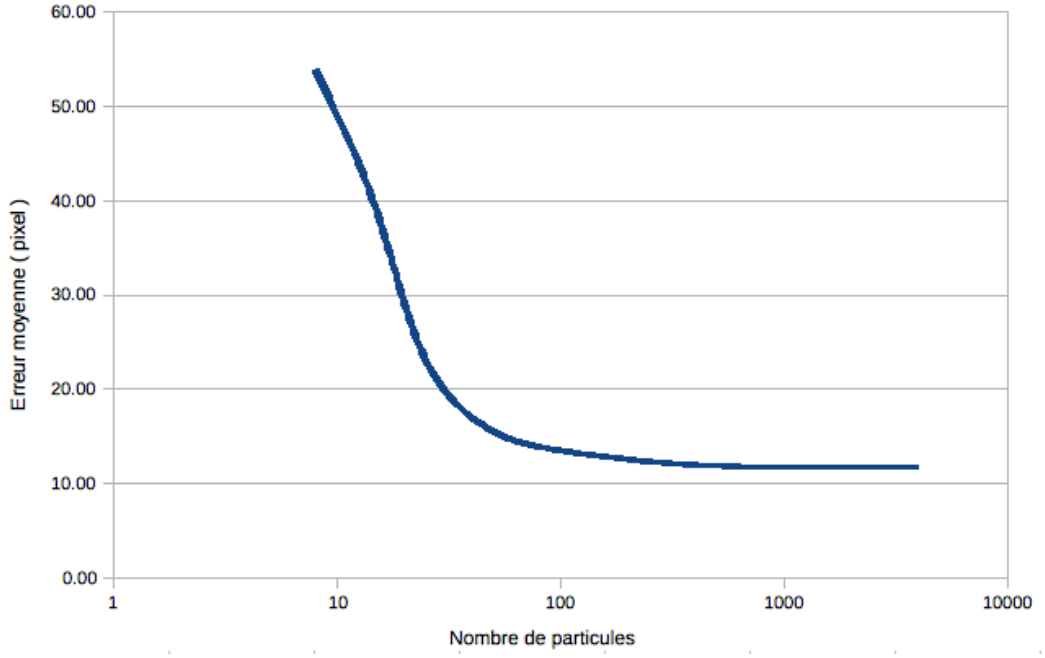


FIGURE 6.34 – Influence du nombre de particules sur l'erreur moyenne de suivi

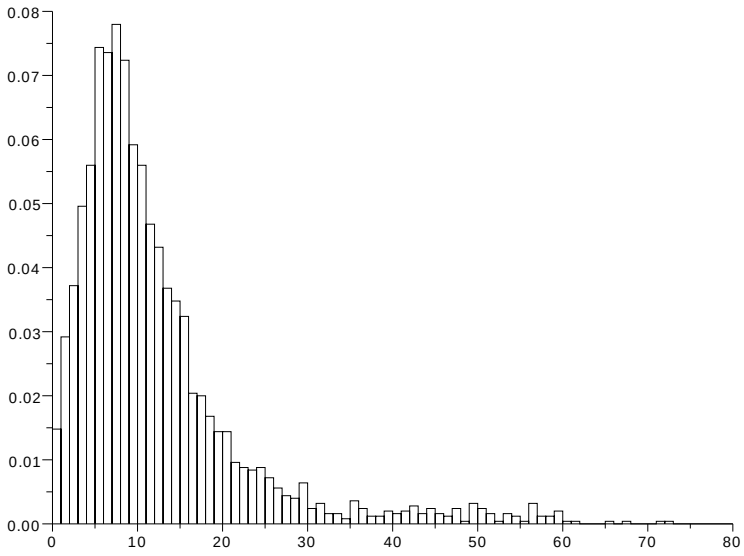


FIGURE 6.35 – Répartition des erreurs de suivi de la main droite (en pixel)

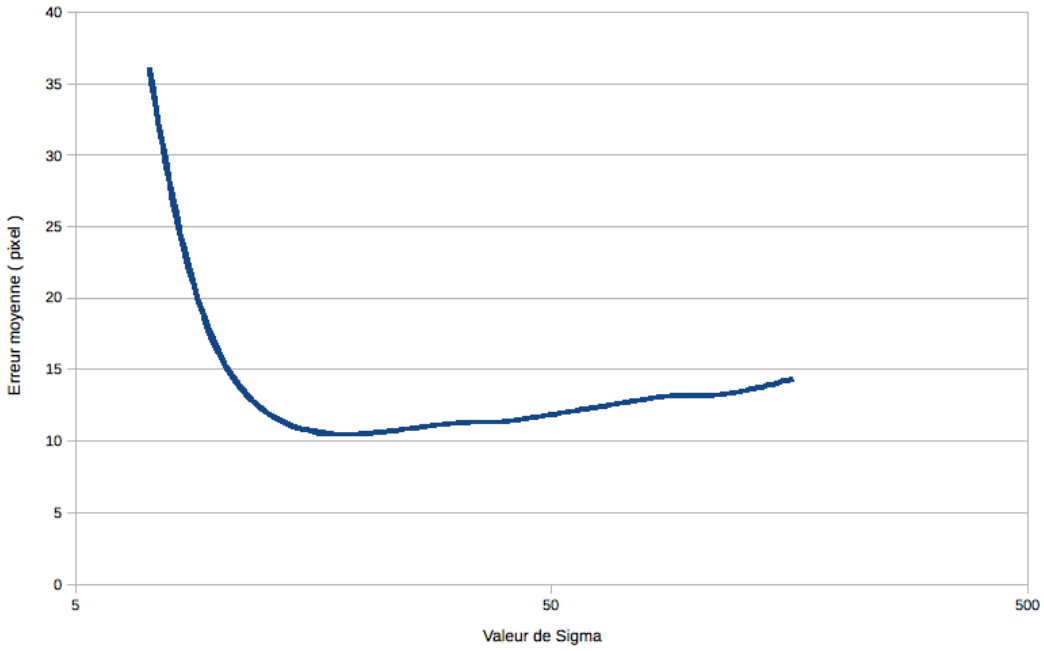


FIGURE 6.36 – Influence de σ sur l'erreur moyenne de suivi

D'après les graphes, la précision optimale du filtre est avoisinée pour un nombre de particules de 1000 et une sélectivité des filtres de $\sigma = 23$. Nous utiliserons ces réglages dans le reste de notre étude.

6.10.5 Précision de l'estimation de la profondeur

L'évaluation de notre méthode d'estimation de la profondeur est basée sur un conte de 7'15" capté simultanément par une caméra de face et une capture de mouvement. La comparaison des profondeurs estimées par cinématique inverse à la vérité terrain fournie par les capteurs magnétiques met en évidence une erreur moyenne de 4 cm, ce qui est suffisant pour une utilisation dans le cadre de reconnaissance automatique de signes, mais trop imprécis pour une reconstruction précise de la posture du signeur. Plusieurs facteurs peuvent expliquer cette erreur d'estimation :

- L'erreur sur l'estimation de la profondeur de la main résulte du cumul des erreurs de suivi 2D.
- La position de la main est en fait celle du centroïde de la zone de peau de la main, visible sur la vidéo. L'estimation de la position de la main est ainsi faussée par la longueur variable des manches du signeur et par les changements de configurations des mains.

Les différents traitements que nous appliquons ensuite à l'estimation de la profondeur permettent ensuite de réduire l'erreur d'estimation. Pour la main dominée, nous notons une amélioration de 15%

en moyenne sur la précision de la profondeur (67% de l'amélioration provient de l'utilisation du lisseur de Kalman et 33% de l'utilisation de la corrélation entre les profondeurs des deux mains).

Ces améliorations prouvent que la prise en compte des spécificités des mouvements des LS dans un processus de correction des données de suivi permet une amélioration de l'estimation de la posture du signeur. Il serait maintenant envisageable d'aller plus loin en effectuant, comme en 8.2.3, une correction de la profondeur orientée par la catégorie des mouvements identifiés. Ce point constitue une des perspectives de notre travail.

6.11 Conclusion

Nous avons présenté dans ce chapitre un nouvel algorithme de suivi de la posture du signeur.
La modification de l'algorithme pour suivre des signeurs à manches courtes n'a pas été évaluée quantitativement, faute de vérité terrain disponible. Cependant, nous livrons nos observations qualitatives sur les résultats obtenus :

- Le filtre ne donne des résultats satisfaisants que lorsque les extrémités des manches sont plus proches des coudes que des épaules du signeur.
- Le résultat du suivi est beaucoup plus dépendant de l'image de peau qui doit intégrer la couleur des avants-bras, de la tête et des mains.
- Le filtre est beaucoup moins précis en cas d'occultation de la tête par les mains.

Certaines informations comme l'orientation de la tête ne sont pas encore déterminées à partir de l'image. La structure de l'algorithme permet néanmoins d'inclure des traitements spécifiques pour les déterminer. Le temps de calcul nécessaire à notre algorithme étant extrêmement faible, des étapes de calcul supplémentaires peuvent être ajoutées sans remettre en cause le fait qu'il puisse traiter les vidéos en temps réel.

Quelques pistes peuvent être évoquées pour améliorer encore la précision et la robustesse de notre algorithme de suivi :

- Une modélisation plus précise de la main de manière à localiser des parties spécifiques de la main plutôt que de localiser un centroïde de zone de peau,
- Une initialisation automatique et optimale de la morphologie du signeur à partir de la vidéo,
- L'utilisation d'informations de texture pour effectuer un suivi plus précis en cas d'occultations,

- La prise en compte d'autres informations comme la forme des mains pour effectuer la désambigüisation.

Nous verrons dans la section 8.2.3, que d'autres informations de plus haut niveau peuvent également être utilisées pour améliorer le suivi. Cela nécessite des modèles de mouvements et de signes que nous présentons dans la section suivante.

CHAPITRE 7

MODELE PARAMETRIQUE DE SIGNE POUR L'ANALYSE DE VIDÉOS

Nous avons décrit dans la section précédente, un algorithme permettant d'effectuer le suivi des mains, des coudes et de la tête d'un signeur dans une vidéo. Nous disposons donc des mouvements absolus et relatifs de chacun de ses membres. Il s'agit maintenant d'exploiter ces mouvements pour en faire une interprétation de plus haut niveau. Nous reviendrons d'abord sur l'état de l'art du chapitre 5 en soulignant plusieurs manques des modèles de mouvement utilisés dans les méthodes les plus citées dans la littérature et nous justifierons le choix que nous avons fait de nous tourner vers des modèles de mouvement plus paramétriques. Notre analyse portera principalement sur la réalisation des mouvements impliqués dans les signes.

Nous proposerons ensuite une classification des signes standards en un certain nombre de catégories de mouvements. Cette classification reposera sur différents travaux déjà effectués dans le cadre de la linguistique, de la génération automatique de Langue des Signes par signeur virtuel, ou de l'analyse automatique de Langues des Signes. Nous examinerons de façon plus détaillée la manière dont il serait possible de décrire de manière paramétrique les mouvements balistiques et les mouvements balistiques répétés. Nous nous attarderons sur la dépendance entre les mouvements des deux mains et les différentes relations de symétries qu'il est possible de mettre en évidence. Nous discuterons également de la projection des mouvements dans l'espace 2D de la vidéo, et ses conséquences sur la caractérisation du mouvement 2D.

À partir des différentes classes de signes et de paramètres que nous aurons dégagées, nous proposerons une méthode pour estimer une similarité entre deux signes. Cette méthode sera utilisée pour effectuer une recherche de signe dans une vidéo en LSF.
Nous achèverons notre discussion sur les modèles de signes en montrant comment des paramètres portant sur d'autres articulateurs (configurations manuelles, rotations du poignet, coudes, labialisation ...) pourraient être pris en compte dans notre méthode de modélisation du signe.

7.1 Retour sur les systèmes de reconnaissance de la LSF

Comme nous l'avons vu dans le chapitre 5, les méthodes les plus utilisées actuellement pour l'analyse de signes sont basées sur les Modèles de Markov Cachés et les Déformations Temporelles Dynamiques. Ces modèles sont basés sur l'hypothèse sous-jacente qu'un signe peut être décrit comme une succession d'états.
Le problème de la variabilité des signes est vite apparu lors de la mise en œuvre de telles méthodes,

si bien que plusieurs modifications ont été apportées pour prendre en compte la variabilité globale du signe dans chacun des états :

- Le découplage de l'analyse des différents articulateurs par le biais de Modèles de Markov Cachés Parallèles,
- L'ajout de cycles dans les chaînes de Markov pour prendre en compte les répétitions,
- La normalisation du signe à reconnaître par rapport à la posture de départ du signe,
- La normalisation du signe à reconnaître prenant en compte son amplitude et son orientation,
- La normalisation par projection du signe sur le meilleur plan.

Ces adaptations ont permis d'obtenir des résultats meilleurs que ceux obtenus avec les HMM et DTW traditionnels, mais font apparaître plusieurs problèmes nouveaux.
Le premier problème, peu mentionné dans la littérature, est de nature calculatoire. La normalisation d'un signe (projection, changement d'amplitude et d'orientation, relocalisation) doit être effectuée pour chaque segment de vidéo avant de pouvoir appliquer les algorithmes de reconnaissance. La conséquence immédiate est qu'il n'est plus possible d'appliquer de programmation dynamique dans l'implémentation des algorithme de reconnaissance. Par conséquent ces algorithmes sont extrêmement lents à exécuter [WB99].
Le second problème tient au découplage des différents paramètres. Il est pertinent d'un point de vu calculatoire comme le montre [VM97], mais ce découplage nie les relations temporelles et géométriques entre les différentes trajectoires. Nous savons pourtant grâce à des études comme [Fil08] qu'un signe peut être modélisé comme des mouvements d'articulateurs, liés par un certain nombre de contraintes qui font partie intégrante de la définition du signe.
Le troisième problème concerne le principe même de l'utilisation de modèles de signes utilisant n'importe quelle déformation temporelle du signe. Ces modèles sont contradictoires avec les observations des linguistes qui soulignent une surprenante régularité dans l'effectuation des signes [Uye97]. D'autre part, nous savons que le même signe effectué avec deux dynamiques différentes peut avoir des sens assez différents.

Les raisons qui précèdent nous ont amenés à envisager une modélisation paramétrique du signe. Les paramètres qui caractérisent le signe peuvent porter aussi bien sur les relations entre les mains droite et gauche, sur la structure spatiale du signe et sur sa structure temporelle.
Nous savons que le calcul des paramètres devra être fait sur chaque segment de la vidéo analysée et que chaque type de mouvement exigera peut-être un mode de calcul des paramètres différent. Cela est envisageable seulement si le nombre de catégories est limité. Pour cette raison, nous centrons dans un premiers temps, notre analyse sur une classification des différents mouvements de la LSF.

7.2 Catégories de signes en LSF

Quelques études comme [Bra96] et [Los00] [Leb98] ont déjà tenté de créer une classification des différents signes de la LSF. Nous présentons sous forme d'un tableau récapitulatif 7.2, les statistiques présentées dans [Bra96] car les catégories utilisées sont relativement proche des nôtres. Les catégories seront illustrées figures 7.3, 7.4 et 7.5. Notre classification des mouvements impliqués

classe	%
droite	42,9
arc	26,1
statique	17,3
cercle	10,9
complexe	2,8

FIGURE 7.1 – Statistiques présentées dans [Bra96] effectuées sur les dictionnaires de [Moo86]

dans les signes se distingue des précédentes, en ce sens qu'elle ne porte que sur la projection 2D des mouvements dans le plan image[1]. Les différentes classes ont été établies à partir des 3243 signes de [MVGD97a] en écartant toutefois les signes composés. Lors de l'analyse des mouvements, il nous est apparu que tous les mouvements balistiques étaient susceptibles d'être légèrement courbés. La distinction entre les primitives "droite" et "arc" ne nous a donc pas semblé pertinente. Par contre, nous choisissons de distinguer les mouvements simples des mouvements répétés (figurés par le symbole "+"), car ils ont une structure spatio-temporelle assez différente. Nous avons également classé les signes par type de relation entre les deux mains.
Les résultats de la classification sont exposés dans le tableau de la figure 7.2. Chacune des classes de mouvement est illustrée par une image de signe.

Nous avons fait le choix de ne pas mentionner dans notre tableau des primitives plus rares qui sont souvent issues de la réalisation simultanée de plusieurs autres primitives :

- Le mouvement balistique répété translaté (fig. 7.4) : 73 réalisations dans le dictionnaire [MVGD97a],
- Le mouvement en croix (fig. 7.5) : 23 réalisations dans le dictionnaire [MVGD97a]

La catégorie de signe statique a été rajoutée artificiellement. Elle ne fait pas uniquement référence aux signes dans lesquels une même posture est tenue durant toute la réalisation du signe, mais plutôt aux signes dont la projection du centre de gravité de la main est quasiment statique sur la vidéo. Cette catégorie contient donc tous les signes impliquant uniquement un changement de configuration et /

[1] Nous faisons ce choix afin de développer des modèles 2D qui puissent tirer parti des résultats de notre suivi qui est beaucoup plus précis dans le plan image.

	Mouvements								
	bal.	bal+	A/R	angle	cercle	cercle+	autre	total	somme
1 main	869	510	42	32	26	180	165	1824	1824
s. centrale	20	25	4	0	0	13	0	62	1030
s. sagittale	357	170	24	13	0	42	25	631	
translation	113	60	4	0	1	12	12	202	
alterné	5	84	3	0	0	43	0	135	
statique	0	0	0	0	0	0	334	334	389
autre	16	4	1	2	0	0	32	55	
total	1380	853	78	47	27	290	568	3243	3243
somme	2233		78	47	317		568	3243	3243

FIGURE 7.2 – Statistiques effectuées par notre équipe à l'aide de [MVGD97a]

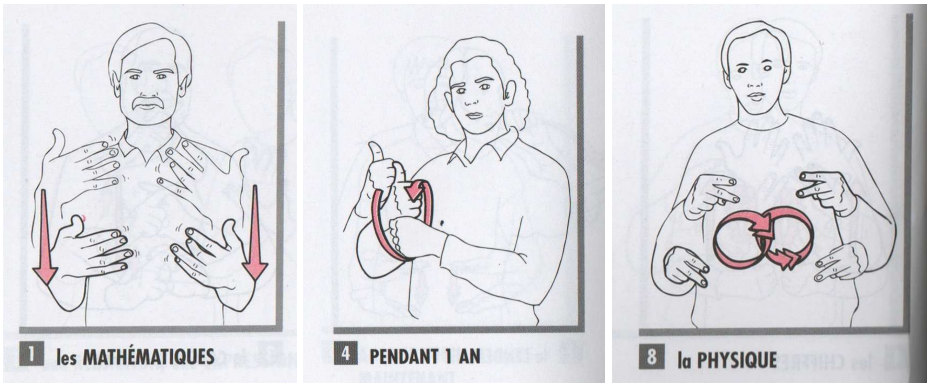


FIGURE 7.3 – Exemple de mouvement balistique [MATHEMATIQUES], circulaire [PENDANT 1 AN] et circulaire répété [PHYSIQUE]

ou d'orientation.

En prenant en compte chacune des catégories énumérées dans le tableau et en incluant la catégorie main statique, nous sommes en mesure de caractériser plus de 92% des signes répertorié dans le dictionnaire [MVGD97a].
Soulignons une nouvelle fois que la classification de mouvements proposée ne tient compte que de la projection 2D des trajectoires des mains. Nous ne considérons pas d'autres traits distinctifs des mouvements comme la profondeur, l'orientation, la configuration et les autres paramètres non manuels.

7.3 Modélisation de la variabilité dans l'exécution des mouvements

Nous avons décidé de centrer notre analyse sur la paramétrisation des mouvements balistiques et des mouvements balistiques répétés car ces deux types de mouvements sont les plus représentés en LSF

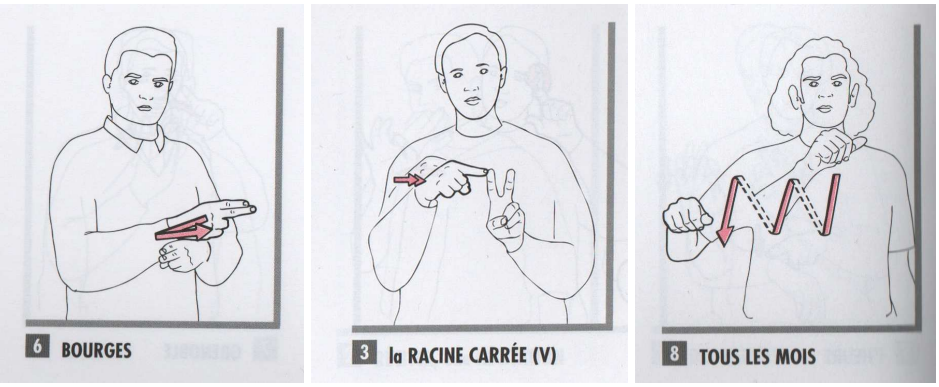


FIGURE 7.4 – Exemple de mouvement d'aller-retour [BOURGES], balistique répété [RACINE CARRE] et balistique répété translaté [TOUS LES MOIS]

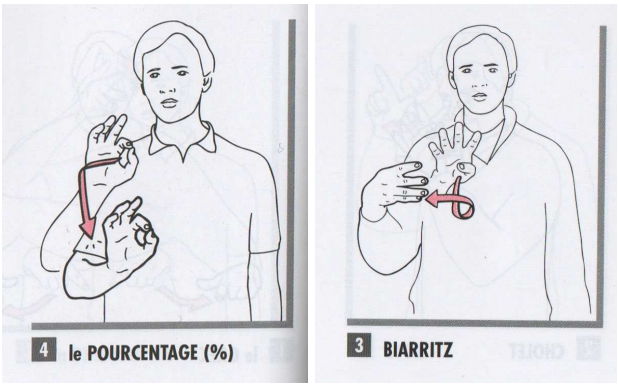


FIGURE 7.5 – Exemple de mouvement en angle [POURCENTAGE] et en croix [BIARRITZ]

et regroupent à eux seuls plus de 69% du lexique auquel nous nous sommes intéressés[2].

Nous pouvons, d'après la littérature, isoler plusieurs paramètres[3] relatifs à la structure spatiale et la structure temporelle du signe. Nous définissons chacun de ces paramètres dans l'énumération qui suit. Le sens précis de ces paramètres sera ensuite défini dans la suite de ce chapitre. Dans l'énumération qui suit, nous associons à chaque fois entre parenthèse le nombre de valeurs qui le quantifient. Les paramètres ne s'appliquant qu'aux mouvements répétés sont marqués du signe "*". Nous ne prétendons en aucun cas que la liste de paramètres présentés soit exhaustive mais elle permet d'expliquer la majorité des variations que nous avons constatées sur les différents types de mouvements. Il manque en particulier tous les paramètres relatifs à la variation de profondeur des mains et ceux permettant de caractériser les mouvements secondaires. Les différents paramètres que nous utilisons seront explicités un à un dans la suite de cette section, en nous basant sur les exemples des mouvements balistiques, avec et sans répétitions.

[2]Ce pourcentage reflète la fréquence des mots des différentes catégories présents dans les dictionnaires. Il est difficile de mesurer la répartition des signes lorsqu'ils sont utilisés en contexte. Les expériences que nous avons menées sur des corpus réels ont toutefois montré une proportion très importante de mouvements balistiques répétés et non répétés.

[3]Le sens que nous donnons ici au mot paramètre est moins restrictif que celui de Stokoe. Il s'agit de toute mesure à même de caractériser le signe.

Les paramètres de la première catégorie affectent la structure spatiale du mouvement :

- Emplacement du mouvement (§7.6.2)
- Orientation du signe (§7.4)
- Position relative des mains (§7.6.2)
- Relation de symétrie (§7.5)
- Courbure de la trajectoire (§7.4)
- Amortissement de l'amplitude* (§7.3.2)

Les paramètres de la seconde catégorie affectent la structure temporelle du signe :

- Régularité (aplatissement du profil de vitesse) (§7.3.1)
- L'asymétrie du profil de vitesse (§7.3.1)
- La durée du signe (§7.3.1)
- Nombre de répétitions* (§7.3.2)
- Diminution de la période* (§7.3.2)
- Durée de la tenue (§7.3.1)

7.3.1 Modélisation des mouvements balistiques

L'analyse de mouvements balistiques est extrêmement importante dans la mesure où ce type de mouvement permet de modéliser plus de 42 % des signes et est à la base de signes comme les pointages, qui ont un rôle déterminant dans la transmission des messages en LSF.
Plusieurs profils de mouvements balistiques ($x = f(t)$, $y = f(t)$ ou $z = f(t)$) capturés dans le cadre du projet SignCom sont présentés dans la figure 7.6. Ils permettent d'avoir une idée des profils de mouvement. Les courbes représentent le déplacement de la main en fonction du temps. Nous avons projeté le mouvement sur l'axe (X,Y ou Z) dont la direction était la plus proche de celle du mouvement. Plusieurs observations émergent de ces profils de mouvement :

- Le mouvement balistique est presque systématiquement précédé d'une phase de retrait où la main va dans le sens opposé du mouvement balistique.

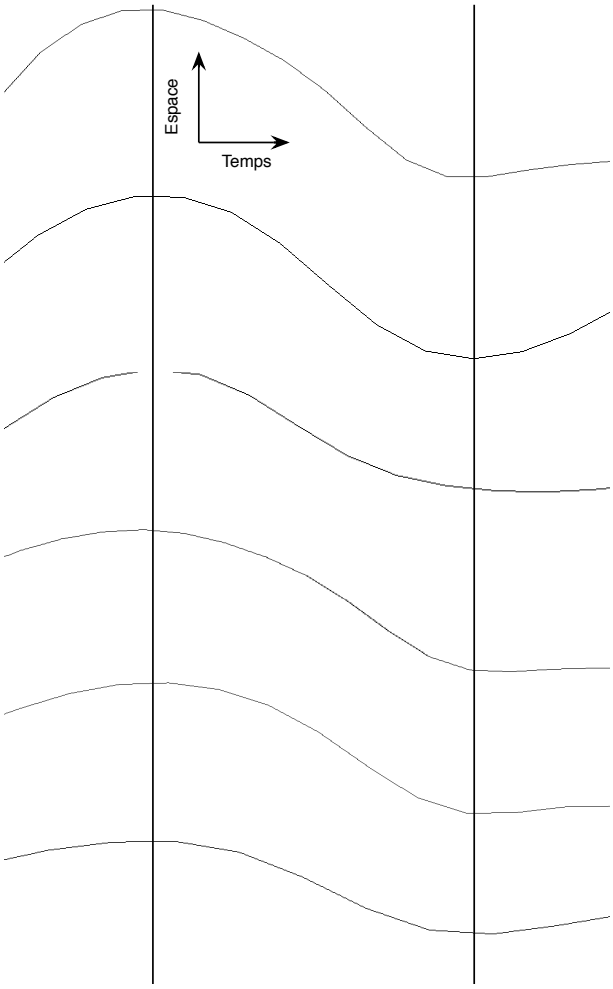


FIGURE 7.6 – Comparaison de plusieurs mouvements balistiques acquis par capture de mouvement.

– Le mouvement balistique est suivi d'une phase de tenue ou d'une phase de retrait.

Nous cherchons maintenant à quantifier les variations des paramètres des différents mouvements balistiques. Nous inspirant de l'analyse proposée par Johnson [LJ90], nous décomposons un énoncé en une succession de phases de mouvements et de phases de tenues en segmentant comme si tous les mouvements avaient la même dynamique. Nous savons d'emblée que les caractéristiques des mouvements ont de grandes chances d'être différentes suivant qu'ils appartiennent à des signes ou à des transitions, cependant le fait d'utiliser des modèles de vitesse identique pour la segmentation nous permettra d'éviter un biais de méthode qui consisterait à utiliser des méthodes différentes de segmentation et à obtenir des différences entre signes et transitions ne résultant que de la méthode de segmentation. Le profil de vitesse utilisé pour les phases de mouvement est représenté sur la figure 7.7 et les tenues sont modélisées par des vitesses nulles. Nous tenons compte de la courbure que peuvent avoir les différents mouvements. Notre modèle de transition permet de modéliser 9 courbures différentes de mouvements balistiques allant d'une trajectoire rectiligne à un demi cercle. Le profil de

vitesse utilisé a été obtenu en moyennant une dizaine de profils de vitesses de mouvements balistiques segmentés à la main. Il correspond à l'équation :

$$v(t) = cos((t - 0.5)\pi) * exp((t - 0.5) * (t - 0.5)) \qquad t \in [-0.25; 1.25]$$

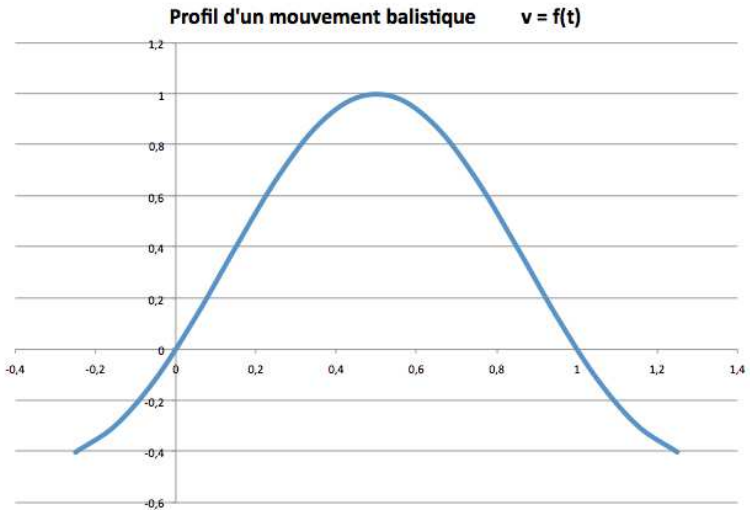


FIGURE 7.7 – Profil de vitesse balistique utilisé pour la segmentation

Nous cherchons à obtenir une régression au sens des moindres carrés entre le profil de vitesse réel de la capture de mouvement et le profil de vitesse théorique obtenu en concaténant les profils théoriques de vitesses correspondant aux phases de mouvements et de tenues. Notons qu'il y a superposition entre les phases de préparation et de retrait des mouvements et les tenues (cf. figure 7.8).

Après avoir vérifié sur la vidéo, que la segmentation proposée était bien cohérente, nous extrayons

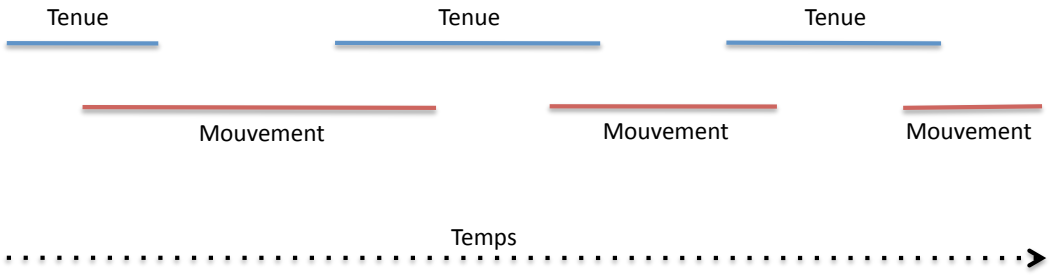


FIGURE 7.8 – Superposition des segments temporels de mouvements et de tenues

des profils de vitesses correspondant à des mouvements balistiques impliqués dans la production de signes et d'autres profils correspondant à des transitions. Pour chaque profil de vitesse, nous extrayons les paramètres suivants :

– L'amplitude du signe A (obtenue par intégration de la vitesse),

– La durée du signe T,

– Le moment où la vitesse est maximale τ/T (cf. figure 7.9),

– La régularité du profil de vitesse σ_v/V_m (cf. figure 7.9),

– La courbure de la trajectoire $c \approx r/D$ (cf. figure 7.10).

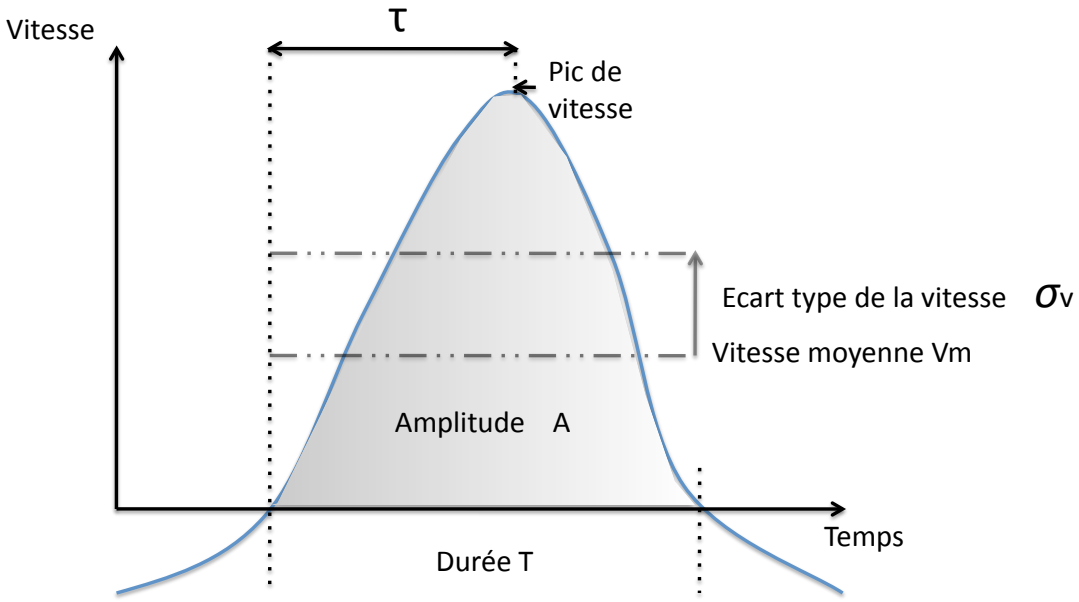


FIGURE 7.9 – Paramètres d'un mouvement balistique

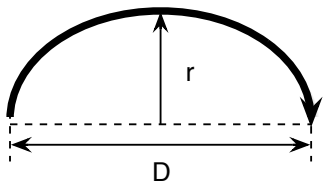


FIGURE 7.10 – Estimation de la courbure $c \approx r/D$ d'un mouvement balistique

Il est possible d'observer sans surprise une augmentation de l'amplitude du signe avec sa durée (dans le cas des signes et des transitions). Par contre, il n'est pas évident de proposer un modèle mathématique satisfaisant expliquant la relation entre ces deux paramètres.
Nous observons également des variations dans l'aplatissement des profils de vitesses. Celles-ci ne

semblent pas corrélées avec les variations d'autres paramètres, et les valeurs ne diffèrent pas entre les signes et les transitions.

Enfin, et c'est là une observation intéressante, nous observons que les pics de vitesses maximales sont situés à un endroit différent pour les signes et pour les transitions. Alors que les profils de transition sont en moyenne symétriques, les profils de signes sont asymétriques et font ressortir une phase d'accélération plus longue et moins brusque que la phase de décélération. Il s'agit selon nous d'un critère phonologique parmi d'autres permettant d'identifier les segments temporels correspondant à des signes. Les profils de vitesses moyens des transitions et des signes balistiques sont représentés figure 7.11 et permettent de constater cette asymétrie. D'un point de vue quantitatif, on trouve des pics de vitesse à $\tau_t/T = 0.50$ pour les transitions[4] et $\tau_s/T = 0.58$ pour les signes balistiques. On note toutefois une variabilité dans la position de ce pic. L'écart type est voisin de $\sigma(\tau/T) = 0.1$ pour les signes comme pour les transitions. L'asymétrie du profil ne doit donc être qu'un indice parmi d'autres pour distinguer avec certitude les signes, des transitions. Par ailleurs, nous avons essayé de

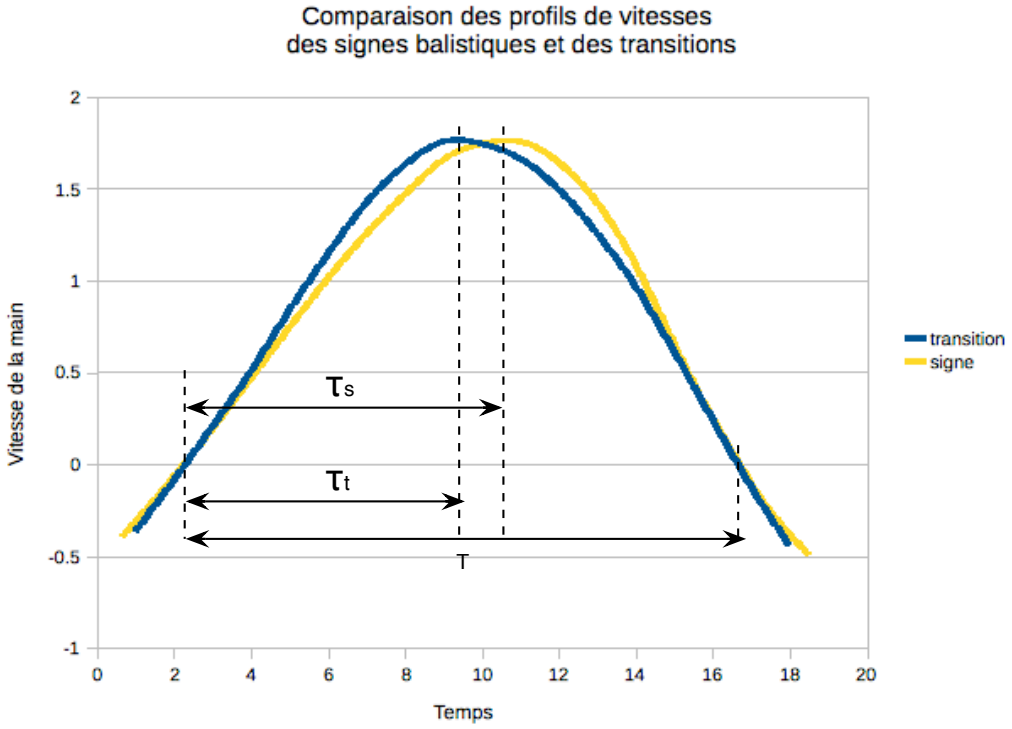


FIGURE 7.11 – Comparaison du profil de vitesse pour un signe balistique et pour une transition

déterminer à l'aide d'une segmentation manuelle une loi mathématique permettant de prédire la durée des phases de tenues en fonction des caractéristiques des mouvements balistiques qui le précédaient. Nous ne trouvons ni de loi, ni de tendance permettant de lier la durée de la tenue aux paramètres du

[4]Ceci va tout à fait dans le sens de la modélisation d'une transition sous forme d'une interpolation symétrique entre un état de départ et un état d'arrivée mais devrait être validé encore par d'autres mesures

mouvement précédent. Ceci peut-être dû à la fois à la faible taille de notre échantillon ou à un défaut de notre méthode de segmentation. Il se peut également que la durée de cette tenue soit régie par des règles qui ne sont pas de nature phonologique ou qu'il faille faire intervenir aussi le mouvement suivant.

Certaines études comme Losson font ressortir des profils de vitesses spécifiques correspondant à des

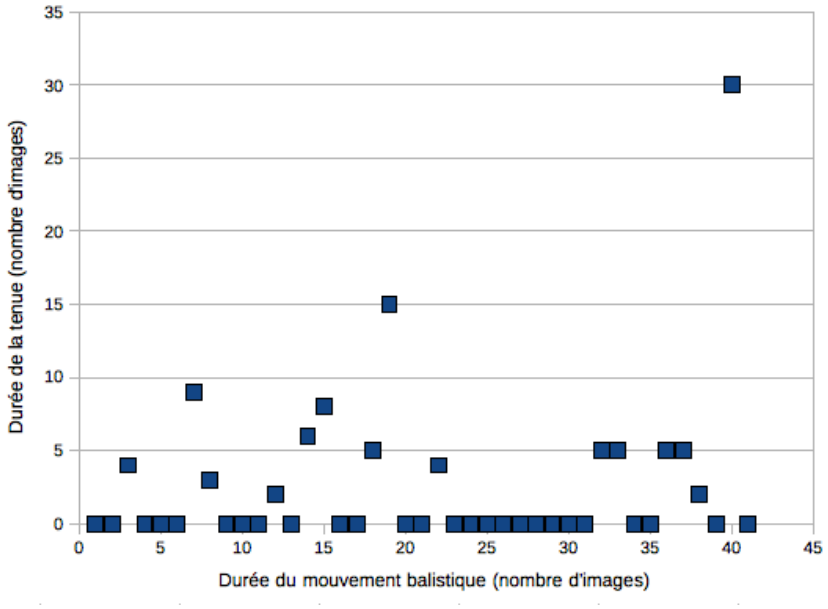


FIGURE 7.12 – Durée de la tenue à la fin des signes balistiques en fonction de la durée du mouvement balistique

mouvements dits "tendus", dans lesquels les phases d'accélération et de décélérations sont brèves et la majorité du signe est effectué à vitesse constante comme le signe [LONGTEMPS]. Nous n'avons pas pu établir une dichotomie claire entre des mouvements "tendus" et des mouvements "non tendus" sur la base du paramètre de régularité de la vitesse du signe. Peut-être était-ce dû au fait que les mouvements "tendus" étaient sous représentés dans notre corpus ?

7.3.2 Modélisation des mouvements balistiques répétés

Le deuxième type de mouvement que nous avons cherché à modéliser est le mouvement balistique répété. Les points de rebroussement qui sont systématiquement présents dans ce type de signes permettent un alignement temporel aisé de différentes instances de répétitions. Nous exposons dans la figure 7.13. plusieurs réalisations de signes à répétition.

Un mouvement balistique répété se présente comme une succession de mouvements balistiques. On distingue sur les exemples de signes à mouvements balistiques répétés de la figure 7.13, que le premier mouvement balistique a une structure différente des autres mouvements balistiques constituant le signe. Son amplitude ainsi que sa durée semblent plus importants. Les mouvements balistiques qui

suivent sont tous d'amplitude et de durée sensiblement égales même si on note de temps en temps un léger amortissement. Comme le mouvement balistique, le mouvement balistique répété peut se terminer par une phase de tenue.

Nous utiliserons, pour mesurer les variations des mouvements balistiques répétés, les notations présentées dans la figure 7.14. Nous faisons le choix de ne pas considérer les amortissements et les légers changements de période entre les différents mouvements balistiques suivant le premier car le taux d'échantillonnage de la capture de mouvement de 30 images par seconde dont nous disposons ne nous permet pas de déterminer avec précision les durées des périodes (de l'ordre de 2 à 3 images).

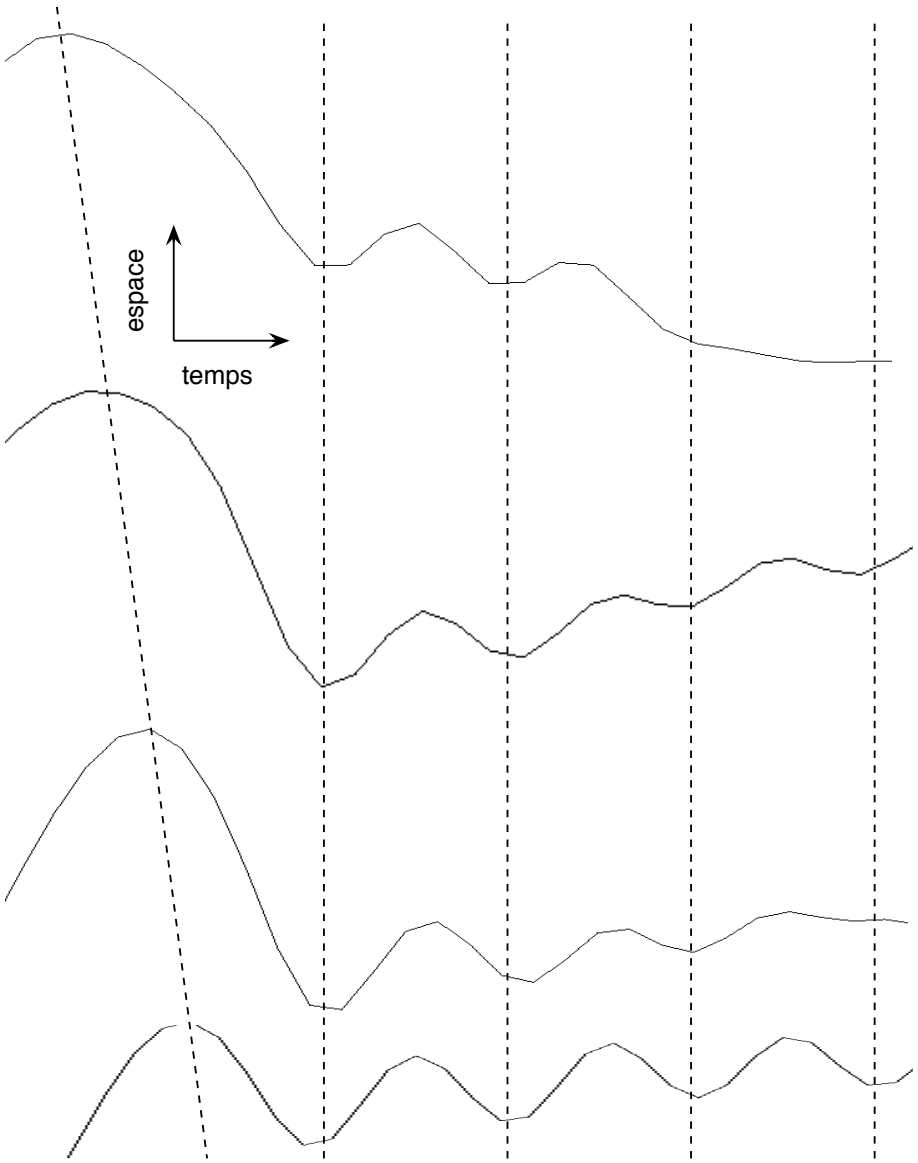


FIGURE 7.13 – Comparaison de plusieurs mouvements balistiques répétés acquis par capture de mouvement pour différents signes

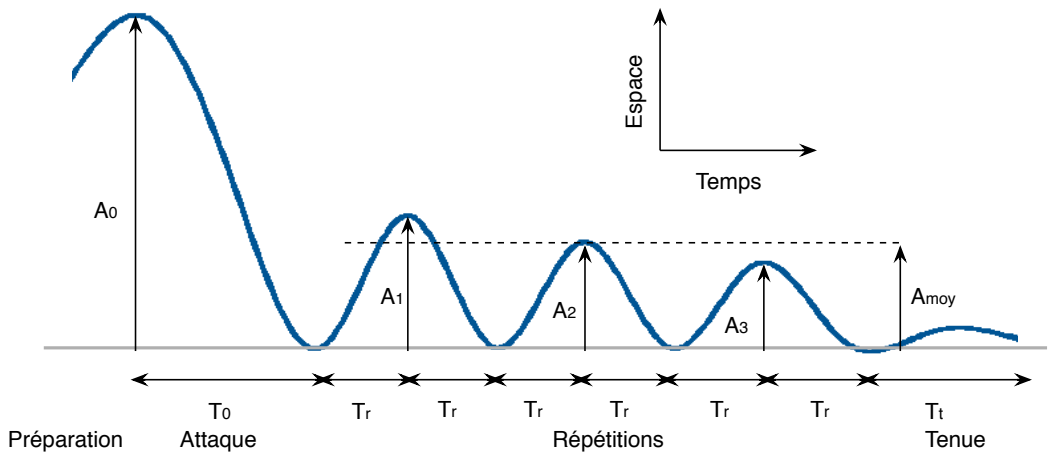


FIGURE 7.14 – Structure du mouvement balistique répété des signes de la LSF

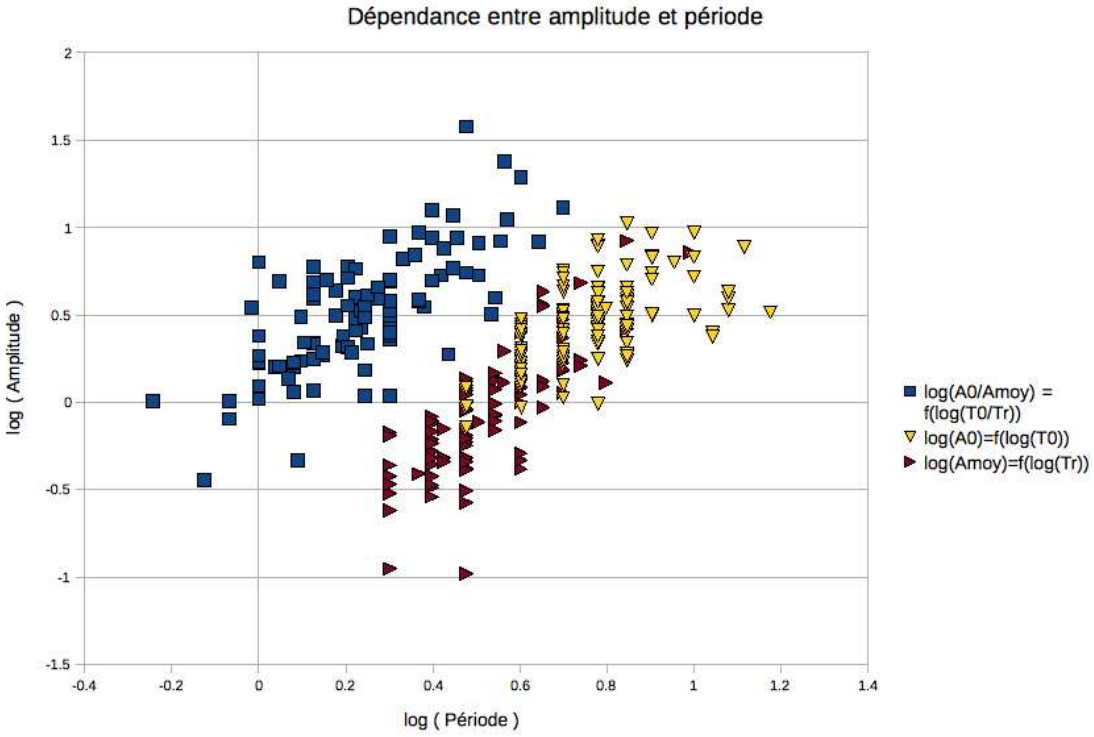


FIGURE 7.15 – Relation entre amplitudes et durées

Nous avons cherché une relation simple entre la diminution d'amplitude et la diminution du temps

(figure 7.15). La relation trouvée est $log(\frac{A_0}{A_1}) \approx \alpha.log(\frac{T_0}{T_r})$. Cette relation de dépendance entre amplitude et durée peut aussi être retrouvée dans chacun des mouvements balistiques puisqu'on note une relation générale $log(A) \approx \alpha.log(T) + \beta$. Nous déterminons une constante $\alpha \approx 1.5$. Cette loi est finalement relativement proche de la loi d'isochronie défini par Freeman [Fre14] cité par [Gib02].

Nous n'avons pas été en mesure de déterminer les asymétries dans le profil des vitesses des différents mouvements balistiques composant le mouvement balistique répété. Il était en effet hasardeux de faire une quelconque hypothèse sur l'allure du profil, alors que les mouvements duraient parfois à peine 3 images.

7.3.3 Modélisation d'autres mouvements

Les études que nous venons de présenter portent chacune sur une centaine de mouvements de signes balistiques, de signes balistiques répétés et de transitions. Nous n'avons pas eu matériellement le temps de mener une analyse aussi approfondie sur toutes les catégories de mouvement les plus fréquentes en LSF. Toutefois, il est utile de montrer que la décomposition du signe en trois phases de préparation, de mouvement principal, et de retrait n'est pas réservée qu'aux mouvements balistiques. Nous exposons dans la figure 7.16 une réalisation d'un mouvement circulaire représentative des réalisations de quelques autres mouvements circulaires que nous avons pu observer. L'orientation de

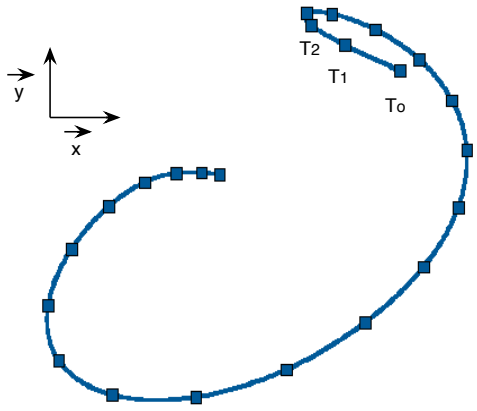


FIGURE 7.16 – Trajectoire type d'un mouvement circulaire réel effectué par un signeur échantillonnée à 30 points par seconde

la vitesse varie quasi-linéairement comme nous pouvons le voir sur la figure 7.17, par contre la vitesse subit de fortes variations. Il serait naturellement nécessaire d'effectuer des mesures d'orientation et de vitesses sur d'autres signes pour voir dans quelle mesure ces observations sont reproductibles. Il est frappant de voir à quel point la trajectoire réelle diffère d'un cercle parfait tel que nous pouvons

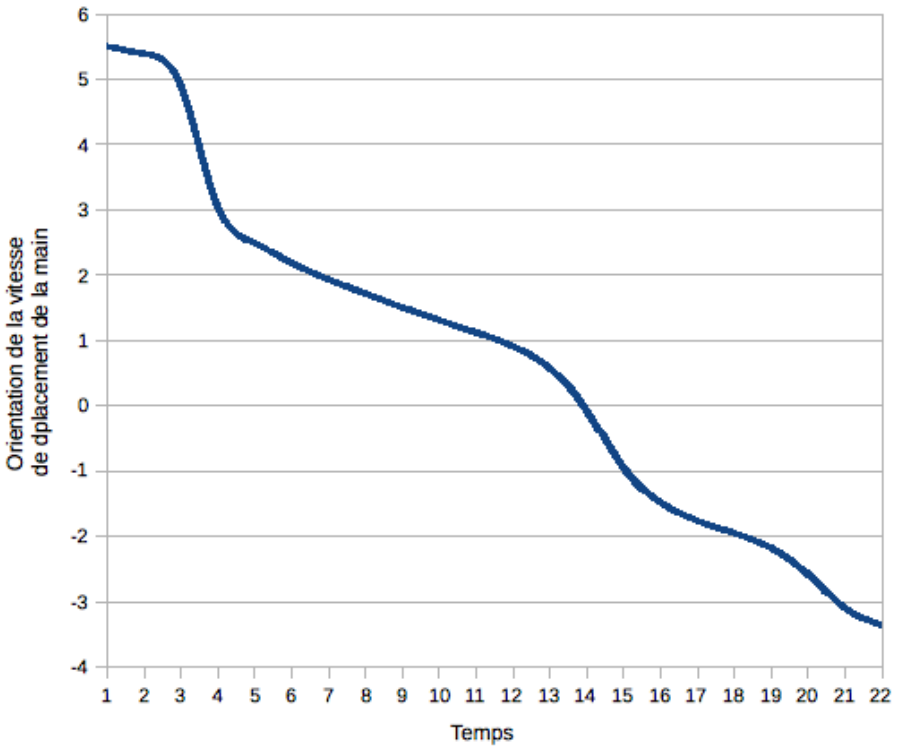


FIGURE 7.17 – Orientation de la vitesse d'un mouvement circulaire réel effectué par un signeur

le voir sur les figurations graphiques de signes. Nous reviendrons sur la différence entre trajectoire réelle et trajectoire idéale dans la partie 8.2 sur les photosignes.

7.4 Utilisation des patrons de mouvements pour la catégorisation de mouvements

Nous venons de présenter différentes variations possibles du mouvement en quantifiant davantage le domaine de variation des mouvements balistiques et des mouvements balistiques répétés qui constituent la base des mouvements de la majorité des signes en LSF. Ces modèles ont un intérêt direct pour la synthèse d'énoncés en LS mais sont aussi utilisables dans un contexte d'analyse.
Nous cherchons à résoudre le problème suivant : étant donné une série de position successives de la main, à quelle catégorie de mouvement appartient son déplacement ?
La première méthode pour résoudre ce problème consiste à effectuer une recherche des paramètres de mouvement permettant d'obtenir un profil de vitesse le plus proche possible du profil de vitesse déterminé par traitement d'image, puis à estimer une distance entre les différents profils de vitesses. Ceci conduirait à résoudre un système d'équations non-linéaires comportant un minimum de 4 inconnues dans le cas de mouvement balistiques simples[5]. C'est ce qui a été fait dans l'étude [WB99]

[5]Dans le cas d'un mouvement balistique, les 4 paramètres en question seraient la courbure, l'aplatissement du profil de vitesse, l'amplitude et l'orientation du mouvement

au prix de coûteux calculs.
La seconde méthode que nous avons envisagée consiste à résoudre ce problème de manière approchée. Nous utilisons pour cela des “patrons de mouvements” qui ne sont autres que des prototypes de mouvements utilisés dans différents types de mouvement. Il suffit ensuite de chercher la dilatation temporelle, la dilatation spatiale et la rotation dans l'espace image qui permettent à ces patrons de mouvement d'être les plus proches possibles des vitesses évaluées à partir du suivi, puis d'estimer la distance entre le profil de vitesse du patron transformé et le profil de vitesse réel.
Les transformations que nous venons d'énumérer (dilatation temporelle, dilatation spatiale et rotation) ne permettent que de prendre en compte la variation d'orientation du signe, de la durée et de l'amplitude. Pour tenir compte de la variation des autres paramètres, il est parfois nécessaire d'utiliser plusieurs patrons de mouvements différents pour modéliser une catégorie de mouvement (par exemple des patrons de mouvements balistiques avec différentes asymétries ou différentes courbures).
Un patron de mouvement est composé d'un patron géométrique et d'un patron de dynamique. Nous définissons ces notions dans les sections qui suivent.

7.4.1 Patrons géométriques

Nous appelons patrons géométrique d'un mouvement la description des différentes orientations (et sens) de vitesses lors de la réalisation du signe. Ce patron géométrique peut être décrit comme une fonction associant un angle à chaque instant. Soit un signe commençant à l'instant t_b et finissant à l'instant t_e, nous appelons t l'instant courant et nous effectuons une normalisation temporelle du signe avec la variable $\tau = (t - t_b)/(t_e - t_b)$. Nous appelons $\alpha(\tau)$ la fonction qui à chaque instant de la réalisation du signe, associe l'angle que fait sa vitesse avec l'axe $\vec{x}$ du repère. Cette orientation ne peut être connue qu'à une constante près que nous nommerons θ (cet angle correspond à la rotation du mouvement réel par rapport au patron de mouvement de référence). La figure 7.18 présente des patrons géométriques associés à différents types de signes que nous avons cherché à décrire.

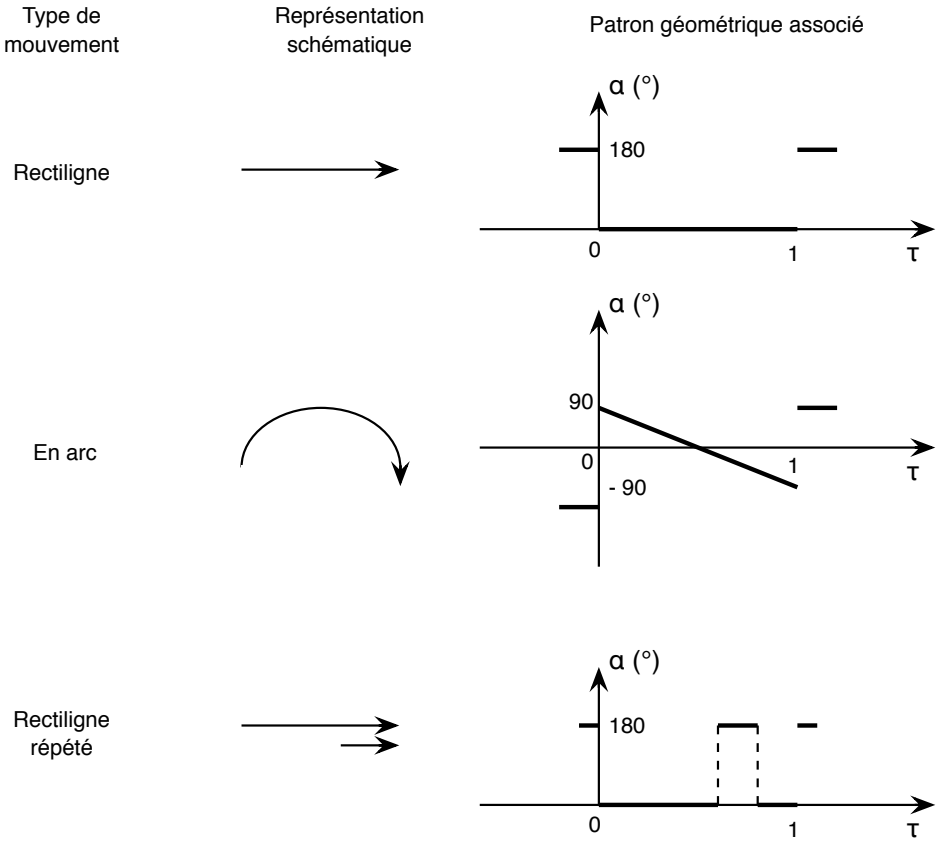


FIGURE 7.18 – Exemples de patrons géométriques

Le lecteur remarquera que les patrons géométriques de signes décrivent également l'orientation des vitesses avant le début et après la fin du signe[6]. Ceci permet de modéliser la préparation et le retrait de la main qui jouxtent le signe proprement dit. Nous notons dans la suite du texte δ_p la durée de la préparation du signe et δ_r la durée du retrait. La durée de la préparation et du retrait est fixée empiriquement à un tiers de celle du signe.

La constante θ représente ce que nous appellerons *l'orientation du mouvement*[7]. Elle peut être déterminée à l'aide du calcul simple qui suit. Soient $v_x(\tau)$ et $v_y(\tau)$ les composantes de la vitesse projetée des mains dans le plan de la vidéo à l'instant t.

$$\overrightarrow{Lin} = \int_{-\delta_p}^{1+\delta_r} \left(v_x(\tau)cos(\alpha(\tau)) + v_y(\tau)sin(\alpha(\tau))\right)\vec{x}$$

$$+ \left(v_y(\tau)cos(\alpha(\tau)) - v_x(\tau)sin(\alpha(\tau))\right)\vec{y}\, d\tau$$

[6]Par début et fin de signe, nous entendons ici, les instants situés au début et à la fin du mouvement où la vitesse de déplacement de la main est minimale. Nous sommes persuadés que les phases de préparation et de retrait ont un intérêt à la fois pour l'analyse et la synthèse, sans pouvoir en fixer pour autant des délimitations temporelles claires.

[7]Attention, la notion d'orientation du mouvement n'est liée d'aucune manière à l'orientation des paumes des mains.

L'orientation du signe est alors l'angle que fait le vecteur $\overrightarrow{Lin}$ avec le vecteur $\overrightarrow{x}$.

$$\theta = (\overrightarrow{x}, \overrightarrow{Lin})$$

Nous pouvons également calculer un score géométrique **GS** qui indique à quel point le mouvement ressemble au patron géométrique.

$$\mathbf{GS} = \frac{|| \overrightarrow{Lin} ||}{\int_{-\delta_p}^{1+\delta_r} \sqrt{v_x(\tau)^2 + v_y(\tau)^2} d\tau}$$

L'illustration 7.19 montre étape par étape comment un mouvement peut être comparé à un patron géométrique afin de calculer son score géométrique et son orientation. Nous nous basons sur l'exemple de mouvements en arc. Une fois l'orientation du signe déterminée, il est possible de projeter les vitesses instantanées sur les directions du patron de mouvement.

$$v'(\tau) = v_x(\tau).cos(\alpha(\tau) - \theta) + v_y(\tau).sin(\alpha(\tau) - \theta)$$

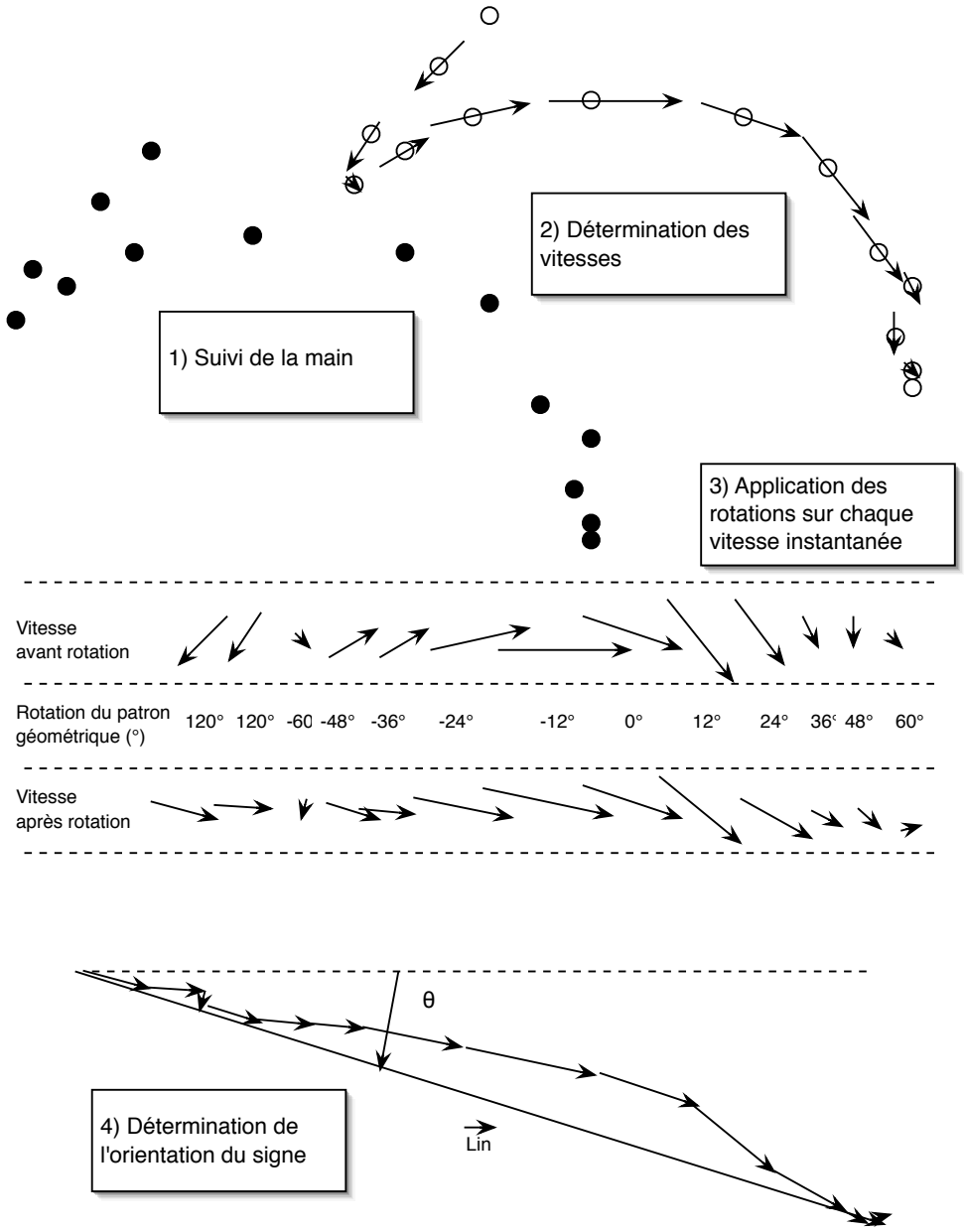


FIGURE 7.19 – Exemple de comparaison d'un profil de vitesse à un patron géométrique

7.4.2 Patrons dynamiques

Nous appelons patron dynamique d'un mouvement, la description du profil de vitesse de signe lors de la réalisation du signe. Nous modélisons ce profil théorique de vitesse par une fonction $v_t(\tau)$. Nous pouvons alors déterminer la similarité entre le profil de vitesse théorique et le profil de vitesse

réel d'un mouvement. Nous appelons **DS** ce score dynamique.

$$\mathbf{DS} = \frac{\int_{-\delta_p}^{1+\delta_r} v'(\tau).v_t(\tau)d\tau}{\sqrt{\int_{-\delta_p}^{1+\delta_r} v'^2(\tau)d\tau}\sqrt{\int_{-\delta_p}^{1+\delta_r} v_t^2(\tau)d\tau}}$$

7.5 Relations main droite/ main gauche

Nous avons jusqu'ici considéré la catégorisation et la caractérisation de mouvements d'une seule main. Nous souhaitons maintenant examiner la relation de symétrie ou de dominance entre les mouvements des deux mains lors de la production de signes.
Plusieurs études linguistiques sur la constitution des signes ont mis en évidence des relation de dominance ou de symétrie entre les deux mains lors de la production d'un signe. Parmi eux [Bat74] émet l'hypothèse que lorsque les deux mains sont mobiles lors de la production d'un signe, leur mouvements sont liés par une relation de symétrie. Bien que cette étude soit appliquée à l'ASL, l'analyse semble transposable à la majorité des signes de la LSF.

7.5.1 Appartenance à un type de symétrie

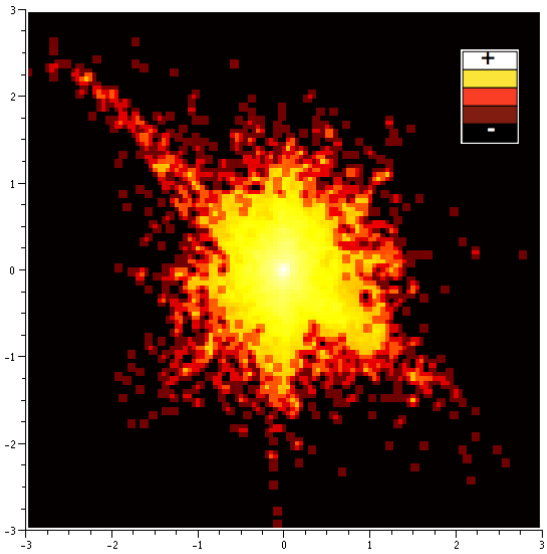


FIGURE 7.20 – Relation entre les abscisses des vitesses des mains droite et gauche

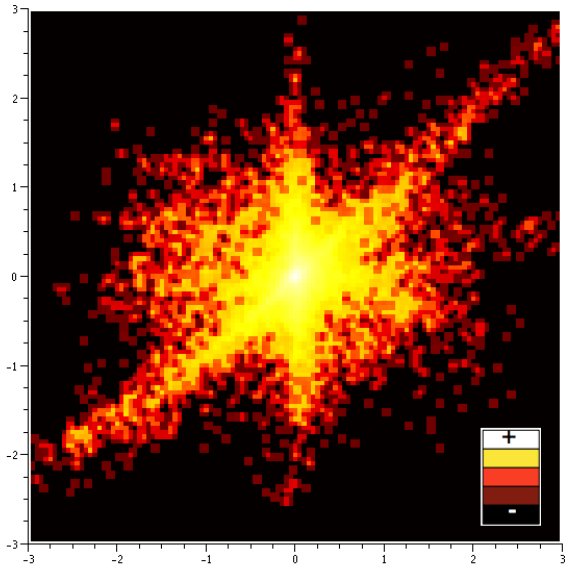


FIGURE 7.21 – Relation entre les ordonnées des vitesses des mains droite et gauche

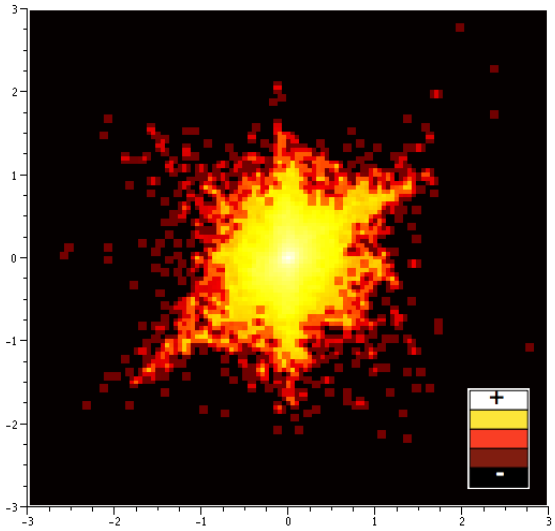


FIGURE 7.22 – Relation entre les profondeurs des vitesses des mains droite et gauche

Comme nous l'avons indiqué précédemment, un signe peut être décrit comme un ensemble de relations entre les différents articulateurs mis en jeux durant la réalisation du signe. Du point de vue du mouvement, ces dépendances se manifestent souvent sous la forme de symétries entre les trajectoires des mains gauche et droite ou de main statiques lors de la production d'un signe.
Ceci ressort nettement lorsqu'on observe dans des données de capture de mouvement les relations entre la vitesse de la main gauche et celle de la main droite. Les graphes des figures 7.20, 7.21 et 7.22 montrent les relations entre les abscisses, les ordonnées et les profondeurs des vitesses des mains

droite et gauche. Les mesures ont été prises sur 30 minutes de vidéo et incluent à la fois les signes et les transitions.

Nous prenons le parti de caractériser les symétries par des relations entre les vitesses des mains droites et gauches, conservées tout au long du signe. Nous utiliserons les dénominations suivantes pour les différentes catégories de symétries en nommant respectivement (vr_x, vr_y) et (vl_x, vl_y) les coordonnées des vitesses des mains droite et gauche :

- Symétrie sagittale : $vr_x \approx -vl_x$ et $vr_y \approx vl_y$
- Symétrie centrale : $vr_x \approx -vl_x$ et $vr_y \approx -vl_y$
- Translation : $vr_x \approx vl_x$ et $vr_y \approx vl_y$
- Symétrie alternée[8] : $vr_x \approx vl_x$ et $vr_y \approx -vl_y$
- Main gauche statique : $vl_x \approx 0$ et $vl_y \approx 0$
- Main gauche statique : $vr_x \approx 0$ et $vr_y \approx 0$

Voici, sur des données réelles, les comparaisons de vitesses que nous obtenons pour des cas de signes à symétrie sagittale et pour les cas de signes dans lesquels la main gauche est statique.

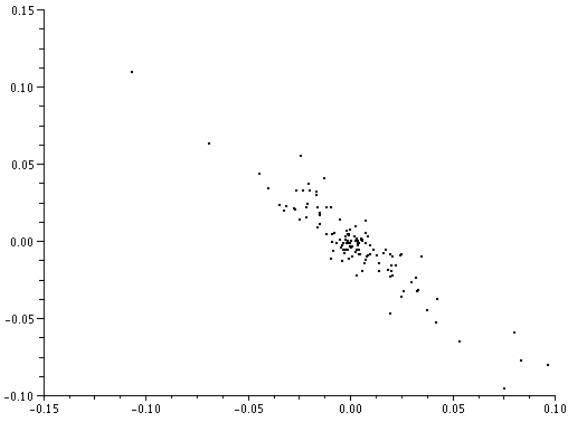


FIGURE 7.23 – Relation entre les abscisses des vitesses des mains droite et gauche lors de la réalisation d'un signe en symétrie sagittale

[8]Nous faisons le choix de traiter les différents signes alternés comme des signes symétriques où $vr_x \approx vl_x$ et $vr_y \approx -vl_y$ bien que cette relation d'égalité entre les profils de vitesse ne soit certainement qu'approximativement vérifiée.

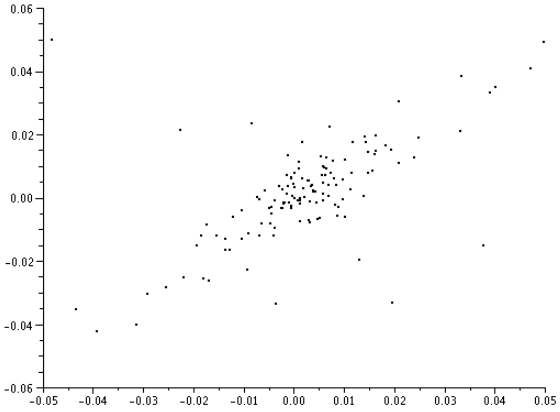


FIGURE 7.24 – Relation entre les ordonnées des vitesses des mains droite et gauche lors de la réalisation d'un signe en symétrie sagittale

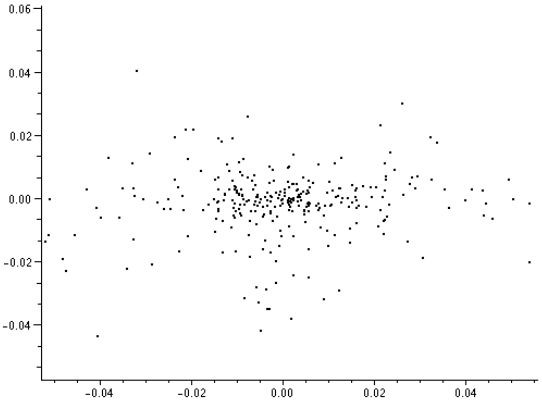


FIGURE 7.25 – Relation entre les abscisses des vitesses des mains droite et gauche lorsque la main droite seule est en mouvement

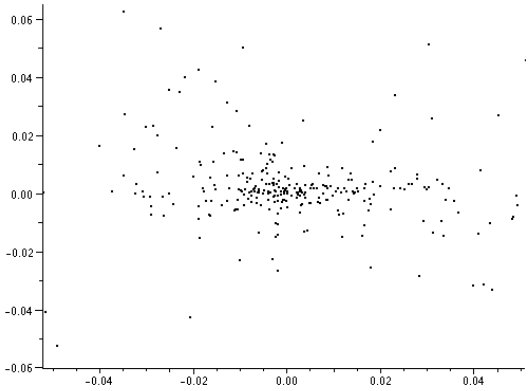


FIGURE 7.26 – Relation entre les ordonnées des vitesses des mains droite et gauche lorsque la main droite seule est en mouvement

A partir de ces statistiques, nous pouvons construire des mesures de similarité **SYM** entre les vitesses réelles des deux mains lors de l'effectuation du signe et les vitesses théoriques qui seraient obtenues si le mouvement était parfaitement symétrique. Nous appelons ($vr_x(\tau)$,$vr_y(\tau)$) et ($vl_x(\tau)$,$vl_y(\tau)$) les coordonnées respectives des mains droites et gauche à l'instant τ et $vr(\tau)$ et $vy(\tau)$ les normes de ces vitesses. Les grandeurs C_x et C_y entre -1 et 1 sont utilisées pour calculer **SYM**. Cette valeur comprise entre 0 et 1 vaut 1 pour un mouvement parfaitement symétrique.

$$C_x = \frac{sgn\left(\int_{-\delta_p}^{1+\delta_r} vr_x(\tau).vl_x(\tau)d\tau\right)\sqrt{\left|\int_{-\delta_p}^{1+\delta_r} vr_x(\tau).vl_x(\tau)d\tau\right|}}{\sqrt{\int_{-\delta_p}^{1+\delta_r} max(vr^2(\tau), vl^2(\tau))d\tau}}$$

La mesure C_y est calculée de la même manière.

$$\textbf{SYM} = \sqrt{C_x^2 + C_y^2}$$

La mesure **MS** permet de détecter une main statique durant l'exécution du signe. Elle vaut 1 si la main gauche est statique et -1 si la main droite est statique.

$$\textbf{MS} = \frac{\int_{-\delta_p}^{1+\delta_r} vr(\tau) - vl(\tau)d\tau}{\int_{-\delta_p}^{1+\delta_r} vr(\tau) + vl(\tau)d\tau}$$

7.5.2 Notion d'angle de symétrie

Nous avons jusqu'ici traité de patrons de mouvements et de symétrie uniquement en deux dimensions. Il est cependant important de garder à l'esprit qu'il ne s'agit que de la projection de gestes tridimensionnels. Cette projection peut engendrer des confusions entre les types de signes suivants :

- Les mouvements balistiques verticaux répétés et les mouvements circulaires dans le plan sagittal,
- Les mouvements balistiques horizontaux répétés et les mouvements circulaires dans le plan horizontal,
- Les mouvements balistiques selon z (profondeur) et les tenues.

Le type de symétrie perçue à partir de la projection 2D du mouvement peut dépendre du point de vue de la caméra et de l'orientation du buste du signeur. Imaginons un signe dans lequel les relations entre les vitesses sont les suivantes :

$$vr_x \approx vl_x \quad avec \quad vr_x \ \ faible$$

$$vr_y \approx -vl_y$$

$$vr_z \approx -vl_z$$

[Théâtre] vue de face
σ = - 125°
Symétrie alternée

[Théâtre] vue de profil
σ = - 71 °
Symétrie centrale

FIGURE 7.27 – Biais de projection

Le signe vu de face sera classé comme un signe à symétrie sagittale. En revanche, le même signe vu de côté sera considéré comme un signe impliquant une symétrie centrale (cf. figure 7.27).

A ce problème s'ajoute celui de la classification dans une catégorie de symétrie de signes, dont l'une des composantes de la vitesse est nulle, ou quasi nulle. Un signe dont les profils de vitesses seraient tels que $vr_y \approx vl_y$ et $vr_x \approx 0 \approx vl_x$ pourrait être à la fois classé dans la catégorie translation et symétrie sagittale. Le résultat dépendrait alors principalement du bruit de mesure dans l'estimation des vitesses. Cela a été un problème pour de nombreux signes du dictionnaire [MVGD97a]. En cas de doute entre deux catégories, nous avons choisi en priorité les catégories de symétrie alternée et de symétrie sagittale.

Pour résoudre ce problème, nous décidons de représenter le type de symétrie sous forme d'un angle σ. Cet angle peut être déduit à partir des grandeurs C_x et C_y que nous avons calculées en 7.5. L'utilisation de l'angle permet ainsi de préserver une représentation cohérente des différents types de symétrie, même quand le mouvement se trouve simultanément dans plusieurs catégories de mouvement. La figure 7.28 montre graphiquement comment est calculé cet angle σ.

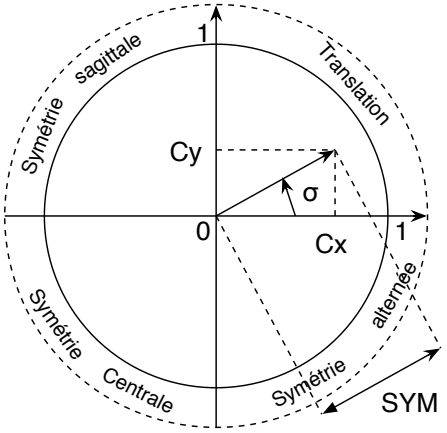


FIGURE 7.28 – Types de symétries

7.6 Mise en œuvre des patrons de signe : Comparaison de deux signes

Les patrons de mouvements et de relation main gauche / main droite que nous venons de décrire permettent d'aboutir à une classification efficace des signes. Nous avons par ailleurs dégagé un certain nombre de paramètres caractérisant les patrons de mouvement. Nous cherchons à nous servir de ces catégories ainsi que des caractéristiques des signes pour estimer une mesure de ressemblance entre deux signes.
Nous évaluons la pertinence des patrons de mouvements dans le cadre de la recherche par le contenu dans des vidéos en LSF. L'utilisateur qui souhaite effectuer une requête de signe effectue un exemple du signe à trouver dans la vidéo. Le programme doit lui présenter les segments de la vidéo qui sont les plus susceptibles de contenir le signe recherché.
Plusieurs étapes sont donc nécessaires :

- Effectuer un suivi dans la vidéo requête et dans la vidéo signe (étape réalisée à l'aide de notre algorithme de suivi).

- Segmenter la vidéo de requête de signe pour délimiter la partie de la vidéo qui correspond au signe recherché.

- Comparer le signe requête à chaque segment de la vidéo cible.
- Classer les segments de la vidéo signe par score de similarité avec le signe requête.

Nous savons que les données de suivi dont nous disposons sont potentiellement entachées d'erreur. Nous savons également que les signes sont susceptibles de varier en emplacement, en orientation, en amplitude. Pour cela, nous écartons d'emblée une recherche basée sur une comparaison directe des trajectoires de signes.
Nous exposons dans les lignes qui suivent une démarche alternative consistant à utiliser les patrons de mouvements pour segmenter le signe requête, puis pour effectuer une recherche des signes qui ont le mouvement le plus ressemblant.

7.6.1 Catégorisation des signes

Nous utilisons la notion de patrons de mouvement pour catégoriser les signes. Les 6 catégories de signes retenues sont les suivantes :

- mouvement balistique simple,
- mouvement balistique répété,
- mouvement angulaire,
- mouvement en aller-retour,
- mouvement circulaire simple,
- mouvement circulaire répété.

Les patrons de mouvements correspondants ont été créés d'après les modélisations de [Los00] en incluant toutefois également les observations sur la structure des mouvements balistiques avec et sans répétition que nous avons mentionnés précédemment comme l'amortissement des mouvements répétés. À chaque catégorie sont associés un patron dynamique et un ou plusieurs patrons géométriques (correspondant par exemple aux différents sens de rotation des cercles)[9]. Nous utilisons les mesures GS et DS définies en 7.4 pour estimer l'appartenance des signes à ces catégories.
Chaque signe implique également une catégorie de symétrie (2 mains mobiles, main gauche statique ou main droite statique). Nous utilisons les mesures MS et SYM pour estimer l'appartenance du signe à ces catégories.

[9]Compte tenu des observations effectuées sur la structure des signes à répétitions, il serait certainement nécessaire d'adopter plusieurs patrons géométriques et dynamique de manière à mieux tenir compte de la variabilité dans la structure des signes à répétition. Nous souhaitons cependant tester dans un premier temps une modélisation des signes la plus simple possible de manière à tester la faisabilité de l'approche.

Nous travaillons cette fois dans un domaine temporel discret. Pour cette raison, les profils d'angles $\alpha(\tau)$ et de vitesses $v_t(\tau)$ sont tous deux discrétisés en N points (nous utilisons N=16 dans notre implémentation).

Nous dénombrons donc en tout $6 * 2 = 12$ catégories de mouvements. L'appartenance d'un signe à une de ces catégories peut être estimée de la manière suivante.

Pour un signe effectué avec la main droite :
$SCORE = GS_r.DS_r.MS$
Pour un signe effectué avec la main gauche :
$SCORE = -GS_l.DS_l.MS$
Pour un signe effectué avec les deux mains :
$SCORE = SYM.\sqrt{GS_r.DS_r.GS_l.DS_l}$

7.6.2 Caractérisation des signes

Les catégories de mouvement que nous venons de mentionner sont discrètes (un signe appartient à une catégorie ou n'y appartient pas). Nous choisissons de caractériser les mouvements à l'aide de valeurs continues. Pour plus de lisibilité, nous faisons le choix de ne pas écrire les formules de calcul associées à l'obtention de chaque valeur et nous décrivons le mode de calcul en français.
Les caractéristiques des différents mouvements sont les suivantes :

L'orientation θ du signe dont le mode de calcul est décrit §7.4,

L'orientation σ de la symétrie dont le mode de calcul est décrit §7.5.2,

L'amplitude A du signe qui est la norme du vecteur $\overrightarrow{Lin}$ dont le mode de calcul est décrit §7.4,

L'emplacement $\overrightarrow{P}$ du signe est relatif à la tête ; il est fourni par le vecteur joignant le barycentre des positions successives de la tête au barycentre des positions successives des mains durant la position des signes,

La position relative $\overrightarrow{R}$ des mains qui est le vecteur joignant le barycentre des positions successives de la main dominante au barycentre des positions successives de la main dominée.

Si un signe est effectué en utilisant la main gauche comme main dominante, la caractérisation du signe sera effectuée en utilisant l'opposé des abscisses dans les données de suivi des mains et de la tête. De cette manière, tout se passera comme si la requête avait été effectuée en utilisant la main droite comme main dominante. Le schéma qui suit illustre les différents paramètres de caractérisations à l'aide d'un exemple de requête de signe.

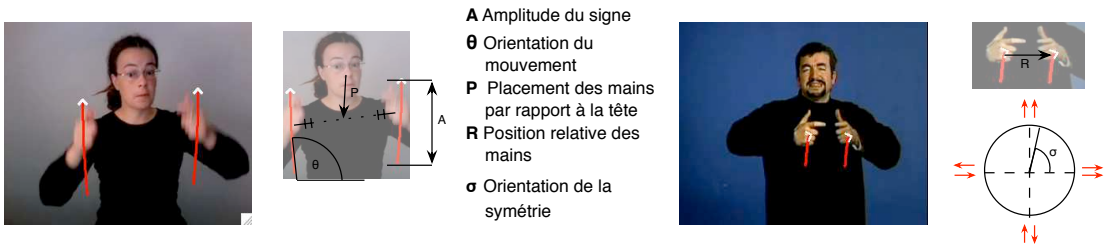


FIGURE 7.29 – Paramètres de caractérisation du mouvement.

Lorsque deux mouvements appartiennent à la même catégorie, il est possible de donner des mesures de similarité partielles pour chacun des paramètres de caractérisation que nous venons d'énumérer. Nous créons donc des opérateurs de comparaison associés à chacun des paramètres renvoyant une valeur entre 0 (les paramètres sont opposés) et 1 (les paramètres sont identiques). Nous présentons ici les opérateurs de comparaison utilisés pour comparer deux mouvements (les indices 1 et 2 permettront de distinguer les mesures concernant chacun des deux signes).

AMP : Opérateur de comparaison d'amplitudes

$$AMP = min(A_1, A_2)/max(A_1, A_2)$$

TS : Opérateur de comparaison d'orientations de symétries

$$TS = 0.5\ cos(\sigma_1 - \sigma_2) + 0.5$$

OR : Opérateur de comparaison d'orientations de mouvement

$$OR = 0.5\ cos(\theta_1 - \theta_2) + 0.5$$

PL : Opérateur de comparaison d'emplacement des signes

$$PL = 0.5\ \frac{\overrightarrow{P_1}.\overrightarrow{P_2}}{max(\overrightarrow{P_1}^2, \overrightarrow{P_2}^2)} + 0.5$$

TR : Opérateur de comparaison d'emplacements relatifs des mains

$$TR = 0.5 \frac{\overrightarrow{R_1}.\overrightarrow{R_2}}{max(\overrightarrow{R_1}^2, \overrightarrow{R_2}^2)} + 0.5$$

7.6.3 Mesure de similarité entre deux signes

Nous cherchons maintenant à estimer la similarité entre un signe 1 dont on connaît la catégorie de mouvement et un signe 2 en nous basant uniquement sur l'information du mouvement. Il s'agit donc de pondérer les différentes mesures d'appartenance du signe 2 à la catégorie du signe 1 et les mesures de similarité basées sur les paramètres de caractérisation du mouvement. La fusion de ces différentes mesures est un point extrêmement délicat sur lequel nous sommes loin d'apporter une solution définitive. Nous pointons dans les lignes qui suivent quelques contraintes que doit respecter une méthode de fusion.

- La méthode de fusion retenue doit être dépendante de la catégorie de mouvement. Il n'est par exemple pas forcément nécessaire de prendre en compte l'orientation des cercles dans le plan image, alors qu'il est fondamental de prendre en compte celle d'un mouvement balistique.

- La méthode doit tenir compte du fait que la main dominante peut être différente pour le signe requête et pour le signe recherché.

- La méthode de fusion devra aussi changer en fonction du nombre de mains impliquées dans le mouvement. Il va de soi que la notion d'orientation de symétrie n'a pas à être prise en compte si une seule main bouge durant la réalisation du signe.

L'idéal serait de bénéficier d'un ensemble d'apprentissage du signe à reconnaître pour effectuer une pondération optimale des différentes mesures d'appartenance aux catégories et mesures de similarité. Nous ne bénéficions malheureusement pas de tels ensembles dans notre exemple d'utilisation. En l'absence de connaissances suffisantes sur la façon dont les signes sont amenés à varier, nous adoptons les méthodes de fusion relativement simplistes. Nous sommes toutefois certains qu'une étude de la variabilité des signes en fonction de leur structure phonologique permettrait de mettre au point des méthodes de fusion plus pertinentes. Pour des raisons de concision, nous ne présenterons que certains cas représentatifs de méthodes de fusion. Les mesures GS_r, DS_r, GS_l, DS_l représentent les scores géométriques et dynamiques des mains droite et gauche.

Méthode d'évaluation de la mesure de similarité entre deux mouvements de rotation effectués à deux mains :

$$SCORESIM = (SYM * GS_r * DS_r * GS_l * DS_l) * (AMP * TR * PL * TS)$$

Méthode d'évaluation de la mesure de similarité entre deux mouvements de rotation effectués avec la main gauche :

$$SCORESIM = (MS * GS_l * DS_l) * (AMP * TR * PL * TS)$$

Méthode d'évaluation de la mesure de similarité entre deux mouvements (hors rotation) effectués avec la main gauche :

$$SCORESIM = (MS * GS_l * DS_l) * (AMP * TR * PL * TS * OR)$$

7.7 Système de requête vidéo

7.7.1 Étapes de la recherche

Nous cherchons maintenant à évaluer la pertinence de notre approche de comparaison de signes dans un système réel. Le problème que nous cherchons à résoudre consiste à trouver dans une vidéo V_t toutes les occurrences $S_t(j)$ d'un signe S_s présent dans une vidéo source V_s.

Plutôt que de résoudre directement ce problème comme dans [Alo06], nous choisissons de le scinder en plusieurs étapes :
A - Caractérisation du signe requête S_s, détermination de sa catégorie cat_s

1. La vidéo requête comportant le signe requête est enregistrée et on effectue le suivi des deux mains du signeur et de sa tête par la méthode décrite dans [LAD09a].
2. Des scores sont calculés pour chaque intervalle temporel $[t_1, t_2]$ et chaque catégorie cat de signe.
3. Les différentes propositions classées par score sont présentées sous forme d'images de signe requête (à la manière de figure 7.27). L'utilisateur peut alors choisir d'après la forme des flèches la catégorie cat_s du signe et la relation entre les deux mains qui correspondent le mieux au signe recherché.

B - Recherche dans la vidéo

1. Un suivi de la tête et des mains est effectué dans la vidéo-cible V_t (cette vidéo n'a subi aucune segmentation ou indexation préalable).
2. Les filtres de la catégorie cat_s et les opérateurs de comparaison sont appliqués sur tous les intervalles temporels de V_t de moins de T_{max} (2 s dans notre implémentation).

3. Les segments de vidéo sont présentés à l'utilisateur en les classant par score de similarité avec le signe source.

L'ensemble du processus est résumé dans la figure 7.30.

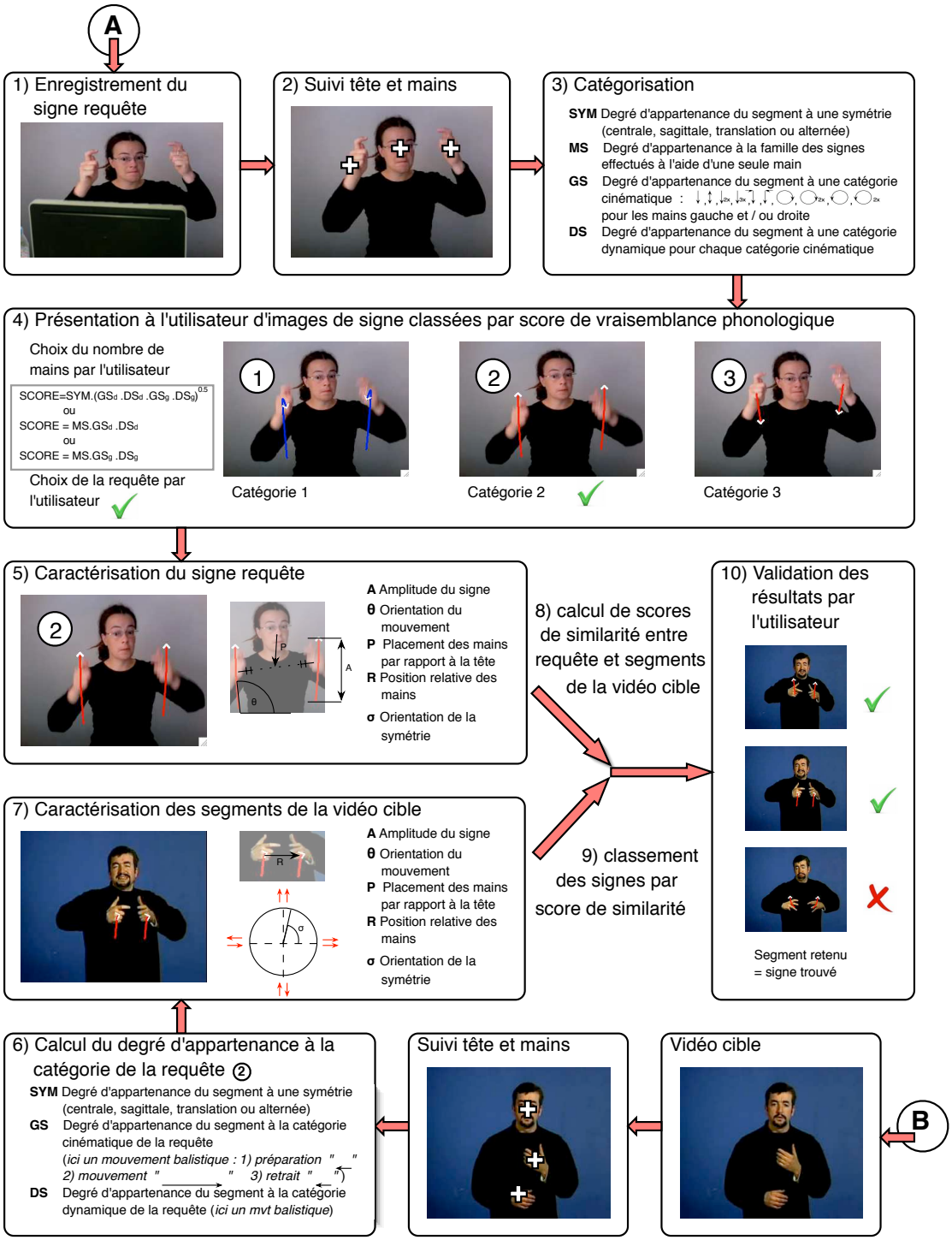


FIGURE 7.30 – Description synthétique du système de requête video

7.7.2 Évaluation du système de requête vidéo

Le but de notre recherche étant de faciliter la requête d'un signe dans une vidéo signée, il est naturel de vérifier dans quelle mesure notre chaîne de traitements permet à l'utilisateur de gagner du temps en ne regardant que les parties de la vidéo cible susceptibles de contenir une occurrence du signe recherché. Pour avoir un panel représentatif des différentes catégories de productions en LSF, les vidéos seront à la fois tirées de récits libres et de traductions de brèves fournies par la société Websourd. L'évaluation a consisté à émettre 103 requêtes différentes. Certains signes étant présentés plusieurs fois, il y avait 178 bonnes réponses possibles. Nous avons volontairement écarté les signes de moins de 3 frames (soit 0.1s) qui sont difficilement identifiables en l'absence de contexte, même par un signeur. Par contre, bien que notre méthode soit basée sur le mouvement, nous n'excluons pas les signes dont la dynamique apparente est faible car ils représentent une partie significative du lexique et peuvent être caractérisés par des critères comme la position relative des mains ou l'emplacement du signe.
Pour chaque requête, notre algorithme extrait chaque segment temporel de la vidéo cible et classe les segments suivant leur taux de similarité avec la requête vidéo. On note à chaque fois le rang relatif de la proposition correcte[10] correspondant à la requête. Le diagramme qui suit présente les effectifs de rangs relatifs obtenus [11].

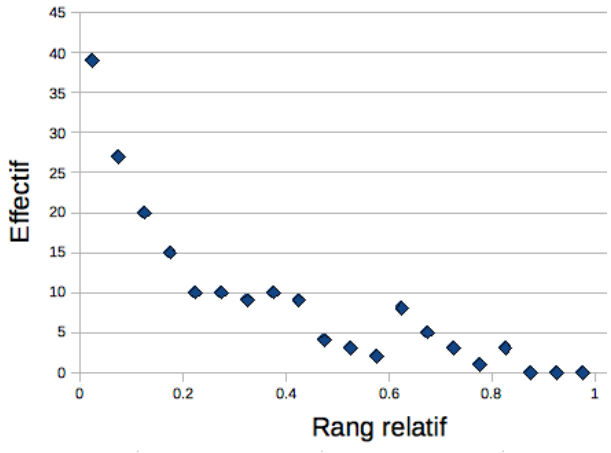


FIGURE 7.31 – Effectif par classe de rang relatif de la bonne réponse à la requête

D'après les résultats, près de 50% des réponses aux requêtes sont présentes dans les premiers segments temporels présentés à l'utilisateur représentant seulement 15% du nombre de segment total. Il est possible de déduire de l'allure de la courbe que la recherche sera environ deux fois plus rapide en visualisant les segments proposés par notre algorithme dans l'ordre de classement.

[10]Le segment extrait de la vidéo est considéré comme correct lorsqu'il contient plus de la moitié d'un signe correspondant à la requête vidéo.
[11]Le rang relatif correspond au rang de la réponse correcte divisé par le nombre de segments de vidéos présentés à l'utilisateur.

Cette moyenne couvre toutefois d'énormes disparités en fonction de la catégorie de signe. Les signes amples comme le signe [IMMEUBLE] présent dix fois dans notre corpus de test peut être localisé en moyenne 15 fois plus vite grâce à notre méthode. Les signes impliquant un mouvement moins ample sont beaucoup plus difficiles à localiser dans la vidéo. On peut émettre l'hypothèse que d'autres paramètres comme l'orientation de la main et la configuration sont dans ce cas nécessaires pour obtenir une caractérisation satisfaisante.

7.7.3 Observations complémentaires

Notre système de requête a été appliqué à des vidéos qui n'étaient pas *a priori* créées pour évaluer un système de requête par le contenu. Ceci est sans doute la raison des résultats assez modestes que nous avons obtenus. Le fait de tester notre algorithme sur des vidéos réelles fait ressortir un certain nombre de problèmes de la recherche par le contenu en LS qui sont rarement mentionnés dans la littérature :

- La structure temporelle des signes peut être amenée à changer. Ainsi, des signes présentés comme sans répétitions dans la requête vidéo peuvent comporter une ou plusieurs répétitions en contexte. Inversement, des signes présentés comme ayant des répétitions peuvent perdre leur répétition. Ce problème avait en particulier été souligné par [Alo06].

- Un signe présenté comme étant effectué à deux mains peut être finalement effectué à une main, soit parce que l'autre main est utilisée dans le cadre d'une structure de grande iconicité, soit par mesure d'économie.

Cela signifie que le signe requête et le signe recherché n'appartiennent pas forcément à la même catégorie. Une étude phonologique sur la variation des différents types de signes serait nécessaire pour permettre de prédire ces variations de catégorie à partir de la catégorie de la requête.

Plus généralement, il serait intéressant de voir s'il est possible de déduire à partir de la structure du signe requête, quels paramètres semblent les plus saillants. Cela permettrait une pondération plus pertinente des différents critères, pour déduire la similarité entre différents mouvements.

7.8 Intégration d'autres paramètres ...

Le processus que nous venons de décrire ne prenait en compte que l'information de mouvement 2D des mains. Il serait dès lors intéressant d'intégrer à moyen terme d'autres paramètres pour estimer plus finement la similarité entre les signes. Ceci nécessiterait toutefois d'améliorer la précision de

l'estimation de la posture et d'adapter légèrement nos modèles de mouvement. Les indices suivants pourraient être utilisés :

La profondeur des mains du signeur peut très facilement être intégrée à notre architecture. Il serait juste nécessaire de modifier légèrement les opérateurs de détection de symétrie et l'utilisation des patrons géométriques. L'utilisation de la profondeur issue de notre reconstruction 3D pourrait être suffisante pour caractériser des signes ou la variation de profondeur est importante comme [BONJOUR] ou [DONNER]. Par contre, le manque de précision rend difficile la caractérisation de mouvements qui n'ont qu'une faible variation de profondeur.

Les coudes pourraient aussi être utilisés dans les patrons de mouvements. On note en effet sur les vidéos qu'ils anticipent de quelques dixièmes de seconde les déplacements de la main. Cette anticipation pourrait être mise à profit dans le cadre de traitement temps réel des vidéos pour prédire les mouvements avant qu'ils soient produits effectivement par la main du signeur.

L'orientation et la configuration des mains pourraient également être utilisées. Etant donné que ces paramètres sont particulièrement difficiles à extraire de la vidéo, il pourrait être intéressant d'utiliser des contraintes que nous connaissons déjà sur le signe pour émettre des contraintes sur les configurations manuelles. À titre d'exemple, nous pouvons repenser aux règles formulées par [Bat74] sur les liens entre les symétries des mouvements et les similarités des configurations manuelles. Il n'est pas certain que l'approche par patron de mouvement soit adaptée aux enchaînements de configurations car le nombre de configurations est potentiellement assez important.

Les articulateurs non-manuels pourraient également être utilisés pour caractériser le signe. Parmi eux, la labialisation et le mouvement de la tête ont déjà été utilisés dans le cadre de reconnaissance de signe à l'aide d'HMM parallèles [MGW00].

7.9 Bilan

Nous avons abordé, dans la section 3.2 consacrée à la modélisation linguistique du signe, les différentes approches linguistiques pour modéliser la structure d'un signe. Certaines approches (multisegmentales) proposent de représenter un signe comme une succession de segments temporels correspondant à des phases de mouvements et de tenues. D'autres (monosegmentales) décrivent les signes comme un seul segment durant lequel les différents paramètres évoluent de manière coordonnée.
L'approche que nous utilisons pour modéliser les signes se rapproche davantage des représentations monosegmentales des signes. Les résultats obtenus montrent que cette approche est pertinente pour la recherche de signe par le contenu.
Nous nous distinguons toutefois des modèles linguistiques en proposant des paramètres continus

pour mesurer les différents paramètres de production du signe (amplitude, orientation, amortissement, courbure ...). Notre démarche globale est finalement assez proche de celle de [Fil08] qui proposait une description de signe sous forme de graphe acyclique dont les arêtes formalisaient des relations entre les différents articulateurs et qui pouvaient être modifiées par des paramètres [FB07].

Notre modèle de signe pour l'analyse mériterait d'être complété avec d'autres types de paramètres permettant de caractériser d'autres articulateurs comme les coudes, les configurations et orientations manuelles. Même s'il n'est pas encore complet, il permet déjà de modéliser de manière satisfaisante la majorité des mouvements, impliqués dans le lexique standard de la LSF. Il est également à même de décrire les mouvements du lexique non standard constitué par les néologismes fréquents en LS qui font généralement appel aux mêmes primitives de mouvements que le signes standards. Par contre, il est probable que la méthode ne soit pas adaptée aux unités gestuelles impliquées dans les transferts de taille et de forme car le mouvement peut traduire la forme de l'objet décrit, et a donc une complexité équivalente à celle de l'objet transféré. Pour une raison similaire, il pourra être délicat d'employer des patrons de mouvements pour caractériser des mouvements d'unités gestuelles impliquées dans des transferts situationnels dans la mesure où les mouvements des mains sont analogues à ceux des entités transférées.

Peut-être serait il possible d'affiner les critères de décision pour permettre d'éliminer les fausses détections de signes. Cela pourrait être effectué avec des méthodes de fusion génériques comme les SVM. Une autre piste d'amélioration serait de modéliser les primitives de mouvements à l'aide de *primitives de mouvements dynamiques* comme c'est le cas dans [MHM10].

CHAPITRE 8

APPLICATION DES PATRONS DE MOUVEMENTS : OUTILS DE TRAITEMENT DE SIGNES

Nous avons exposé dans le chapitre précédent, la notion de patron de mouvement. Notre évaluation a prouvé qu'ils pouvaient être utiles pour effectuer de la recherche par le contenu dans une vidéo en LSF à l'aide d'une requête de signe vidéo. Il existe également de nombreuses autres applications des patrons de mouvement. Dans ce chapitre, nous verrons dans quelle mesure les patrons de mouvement peuvent permettre de résoudre des problématiques liées aux outils de traitement de signes.
Nous définissons d'abord la notion d'outil de traitement de signes par rapport aux outils de traitement de texte existants. Nous montrons ensuite comment utiliser les patrons de mouvement pour générer des images de signes. Enfin, nous discutons sur la possibilité d'utiliser les patrons de signes pour effectuer une segmentation d'une vidéo en LSF. Le chapitre s'achève sur un bilan de ce qu'il est déjà possible de faire dans le domaine du traitement de signes et sur nos apports dans ce domaine.

8.1 Outil de traitement de signes

Nous définissons la notion d'outils de traitement de signes par analogie avec les outils de traitement de texte tels qu'OpenOffice, Word, Pages … Ceci peut sembler paradoxal dans la mesure où les LS n'ont pas d'écriture normalisée. La question de la mise en forme d'une production en LS se pose néanmoins de plus en plus fréquemment lors de la production de documents bilingues, de supports en LS ou de vidéos produites par des signeurs virtuels.
Listons dans un premier temps les différentes fonctions offertes par les éditeurs de texte, avant de les transposer aux LS. Un logiciel de traitement de texte moderne propose les fonctionnalités suivantes[1] :

Créer ou modifier du contenu

- Taper du texte, copier, couper, coller des portions de texte, rechercher et remplacer,

- Insérer des notes de bas de page, entête, pied de page,

- Insérer de tableaux, graphiques, images.

Structurer un document

- Structurer le texte en paragraphes, titres, listes …,

[1] Il ne s'agit que des fonctions les plus courantes, mais la liste n'est pas forcément exhaustive.

- Créer des sommaires ou index,

- Mettre un lien hypertexte,

- Gérer des liens vers des références.

Mettre en forme le document

- Changer la police d'un texte, sa couleur, sa casse, sa taille,

- Créer des types pour les titres, énumération, paragraphes …,

- Optimiser la présentation des documents (coupure des mots au bon endroit, insertion des espaces en respectant la typographie, déformation des mots pour tenir dans des cadres …),

- Mettre en page, Imprimer.

Offrir des fonction de TAL

- Faire des statistiques sur le nombre de caractères, le nombre de mots,

- Corriger la grammaire et orthographe,

- Proposer un dictionnaire des synonymes.

D'autres logiciels spécialisés permettent également d'effectuer de la dictée automatique de texte ou bien de lire le texte par le biais de synthèse vocale.

Une partie de ces fonctionnalités peuvent déjà être remplies par des logiciels de montage de vidéo. Certaines modifications peuvent être effectuées à l'intérieur de la vidéo en LS :

- Couper, copier, coller des fragments de vidéos, puis les monter,

- Insérer des images, des tableaux, du texte, d'autres vidéos à l'intérieur de la vidéo d'origine,

- Structurer la vidéo en chapitres et y insérer des menus et des raccourcis,

- Ajouter des liens hypervidéos (pour les vidéos cliquables).

- Changer l'apparence de la vidéo en jouant sur la vitesse de lecture, la résolution, le contraste, le cadrage,

- Permettre de ré-enregistrer une séquence.

Malheureusement, il se pose plusieurs problèmes difficiles à solutionner. Des problèmes de discontinuité dans l'image apparaissent lors du montage, en raison de la différence de signeur, de la variation de l'habillement ou du changement de posture entre deux images consécutives. Par conséquent, une modification d'un signe à l'intérieur d'un énoncé n'est pas possible, et oblige à un ré-enregistrement de l'intégralité de l'énoncé qui le contient !

Pour résoudre ce problème il est possible de recourir à des signeurs virtuels. Ils offrent différentes fonctionnalités semblables à celles du montage vidéo et permettent en plus, les opérations suivantes :

- Déplacer des signes dans l'espace et corriger des mouvements,
- Remplacer un signe par un autre en assurant la continuité de l'image,
- Concaténer plusieurs signes en assurant la continuité de l'image par interpolation,
- Modifier l'apparence du signeur : sexe, vêtement, morphologie[2] ...
- Intervenir sur la prosodie du signeur en modifiant par exemple, le rythme des signes (time warping) ou l'amplitude des signes.

Suivant la méthode d'animation considérée, deux problèmes de nature assez différente peuvent survenir. Si l'animation est effectuée par capture de mouvement, par rotoscopie, ou par toute autre méthode visant à copier une production de référence, nous manquons actuellement de modèles permettant de déformer l'animation de manière cohérente (pour modifier l'amplitude, l'emplacement, la vitesse d'exécution) dans le respect des contraintes des LS. Si au contraire, l'animation est générée en spécifiant un certain nombre de contraintes aux signeur virtuel (cibles à atteindre, relations entre les mouvements des deux mains, relations entre les positions et les orientations des paumes ...), nous manquons de modèles nous permettant de prédire comment le signe sera effectivement réalisé, car nous savons que les coarticulations, butées articulaires et préférences du signeur vont engendrer un écart par rapport à la description idéale du signe.

En complément du montage vidéo ou de la synthèse par signeur virtuel qui permettent de traiter une vidéo, il est possible d'utiliser d'autres techniques pour composer des documents comportant plusieurs vidéo. Il est ainsi possible d'utiliser la disposition spatiale des vidéo pour faire mieux ressortir la structure logique du document :

- Les documents multimédias comme les pages Web peuvent permettre de présenter plusieurs vidéos, à la suite les unes des autres, pour découper le document en *paragraphes*.

[2]Cela nécessite toutefois des adaptations de cinématique inverse pour garantir le respect des positions relatives des différents articulateurs.

- Les vidéos peuvent être liées à un sommaire textuel qui permet de naviguer plus facilement à l'intérieur des documents en LS. Ceci conduit toutefois à utiliser une autre langue écrite ou des symboles graphiques.
- Pour éviter cet écueil, il est possible de remplacer les titres par de petites vidéos. Ceci est envisageable lorsque le nombre de titre est restreint ; le site www.websourd.org/ montre un bon exemple de vignettes animées. Cette solution est toutefois difficile à mettre en oeuvre lorsque les vidéos de titre sont trop nombreuses. D'une part, les informations ne sont pas persistantes, d'autre part, il est difficile de se concentrer sur une vidéo lorsque les autres sont en mouvement.

De nombreux problèmes relatifs au traitement de signe restent en suspens :

- Même si la modification de signes dans une animation est théoriquement possible, elle reste l'apanage de professionnels formés à l'infographie ou au montage de vidéo. La modification de signes met en œuvre des règles empiriques, souvent difficiles à formuler.
- Il est encore assez difficile de sélectionner un signe pour lui appliquer une modification, cela nécessiterait une segmentation de la vidéo en signes.
- Il manque des outils pour créer des représentations statiques de signes, pouvant être utilisées pour créer des titres ou imprimer un document.
- Il n'existe que peu d'outils de TAL permettant d'assister une personne dans la création de documents en langue des signes.
- De nombreux progrès restent encore à accomplir dans les algorithmes permettant la dictée automatique ou la synthèse d'énoncés en LS.

Nous n'apportons pas de réponse à l'ensemble de ces problématiques, mais nous montrons comment la notion de patron de mouvement peut être utilisée pour apporter de nouvelles fonctionnalités de traitement de signes. Nous avons déjà abordé dans le paragraphe précédent la problématique de recherche de signe dans une vidéo. Les sections qui suivent montrent comment les patrons de signes peuvent également être utiles pour créer des images de signes, puis pour segmenter une vidéo en signes.

8.2 Génération automatique d'une image de signe

8.2.1 Notion de photosigne

Comme nous l'avons souligné en introduction, un des problèmes récurrent dans la production de documents en LS, est de faire apparaître certaines informations de manière statique (en utilisant une

représentation qui ne fait pas intervenir le temps). Ceci s'avère nécessaire pour mettre en évidence des titres, des liens, imprimer un document ou contenu d'une vidéo, sans pour autant recourir à une langue écrite.

L'utilisation d'une image choisie dans la vidéo de réalisation d'un signe est déjà une solution par-

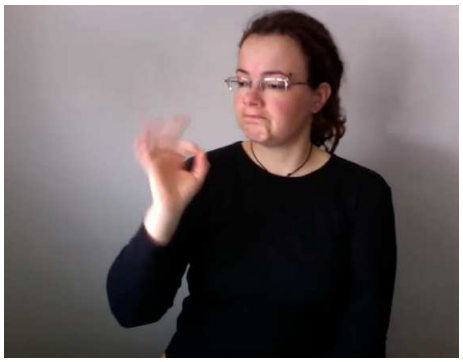


FIGURE 8.1 – Image provenant de la vidéo du signe [OUI].

tielle. Cette image (cf. figure 8.1) permet en effet de restituer la configuration, l'orientation des mains, l'expression du visage et l'emplacement à un instant de la réalisation d'un signe. Le paramètre de mouvement n'est visible qu'indirectement (flou de l'image, position des coudes qui indiquent la dynamique, du mouvement de la tête qui anticipe sur les positions à venir ...). D'autre part, il est fréquent que les signes impliquent plusieurs emplacements et plusieurs orientations.

Comme il n'existe pas d'écriture unifiée des signes, il est nécessaire de mettre au point une représentation 2D du signe permettant de figurer tous les paramètres en utilisant des conventions aisément compréhensibles (par exemple, la flèche figure le mouvement).

Cette approche a été utilisée par les éditeurs de dictionnaires comme Monica Companys [Com06] ou IVT [MVGD97a]. Les résultats obtenus sont très satisfaisants, mais nécessitent un investissement conséquent en temps de travail. A titre indicatif, la réalisation d'une image de signe peu prendre au moins 30 minutes pour une personne non formée et un peu plus de 5 minutes pour une personne formée disposant de logiciels adaptés[3].

Nous avons donc cherché à mettre au point un logiciel pour faciliter la création d'images de signe en automatisant la génération de flèches, et la fusion de plusieurs images prises dans la vidéo du signe (en gérant les transparences). Nous appellerons nos images de signes des "photosignes" et nous exposerons étape par étape la création de telles représentations graphiques de signe.

8.2.2 Reconstruction du mouvement

Un photosigne (cf. figure 8.2) correspond à une, deux ou trois images fusionnées par transparence, auxquelles sont adjointes des flèches figurant les mouvements d'une ou des deux mains. Il est donc

[3]Estimation proposée par une infographiste, qui produit elle même des images de signes à l'aide d'outils d'infographie.

FIGURE 8.2 – Exemple de photosigne correspondant à une version du signe [OUVRIR].

nécessaire, pour créer ces flèches, de connaître précisément les trajectoires des mains.
Nous savons que la reconstruction du mouvement à partir de vidéos monovues peut conduire à des estimations de trajectoires extrêmement bruitées. Cependant nous connaissons également un certain nombre de règles qui peuvent nous permettre de corriger les trajectoires. Nous faisons également intervenir la notion de patrons mouvements dans nos méthodes de correction de la trajectoire.

8.2.3 Nécessité de correction des trajectoires des mains

Il est nécessaire de corriger l'estimation de la position des mains droite et gauche dans l'espace de l'image pour plusieurs raisons :

- Une présence de point aberrant dans une trajectoire rend la synthèse de la flèche inesthétique.

- La personne qui crée le photosigne souhaite visualiser une image du mouvement idéale qui n'est pas affecté par le phénomène de coarticulation.

- Le fait que les flèches représentent une partie des phases de préparation et de retrait des mouvements peut induire en erreur dans la compréhension du mouvement du signe.

Les figures 8.3, 8.4, 8.5 et 8.6 représentent différents problèmes à résoudre pour aboutir à une représentation des signes acceptable visuellement.

FIGURE 8.3 – Photosigne montrant une trajectoire bruitée

FIGURE 8.4 – Photosigne montrant des trajectoires non symétriques à cause de la co-articulation des signes

FIGURE 8.5 – Photosigne incluant une partie du mouvement de préparation

FIGURE 8.6 – Photosigne incluant une partie du mouvement de retrait

Pour aboutir à un résultat satisfaisant, nous mettons en œuvre les étapes de traitement suivantes :

- Catégorisation du mouvement (nombre de main impliquées, type de mouvement, type de relations entre les mains). Pour l'instant cette étape est effectuée par un opérateur humain, mais nous envisageons de l'automatiser à court terme en utilisant les algorithmes de caractérisation décrits dans le §7.
- Correction des trajectoires des mains à partir de la relation de symétrie identifiée.
- Caractérisation des mouvements des mains (amplitude, orientation, courbure ...)
- Générations de flèches représentant le mouvement idéal (que nous savons par ailleurs être différent du mouvement réel que nous avons caractérisé) et esthétique.

FIGURE 8.7 – Etapes de filtrages des données de suivi pour la création d'un photosigne.

Les calculs à effectuer diffèrent pour chaque catégorie de signe. Nous faisons donc le choix d'illustrer uniquement les calculs correspondant à l'exemple de la figure 8.7. Il s'agit d'un signe à symétrie sagittale où les mains décrivent un mouvement balistique en arc.
Le filtrage se déroule en deux étapes :

– Les coordonnées brutes issues de la trajectoire de suivi entre le début du signe $t = t_b$ et la fin du signe $t = t_e$, sont notées respectivement ($x^0r(t)$,$y^0r(t)$) pour la main droite et ($x^0l(t)$,$y^0l(t)$) pour la main gauche. Ces trajectoires brutes sont filtrées de manière à tenir compte de la symétrie du signe. Les coordonnées filtrées sont notées respectivement ($x^1r(t)$,$y^1r(t)$) et ($x^1l(t)$,$y^1l(t)$).

– Un deuxième filtrage est effectué en utilisant la notion de patron de mouvement. A l'issue de ce deuxième filtrage, les coordonnées sont notées respectivement ($x^2r(t)$,$y^2r(t)$) et ($x^2l(t)$,$y^2l(t)$).

8.2.3.1 Utilisation de la notion de symétrie

Dans le cas que nous étudions, le signe implique un mouvement à symétrie sagittale dans lequel nous avons théoriquement les relations $vrx \approx -vlx$ et $vry \approx vly$ (cf. §7.5), d'où la relation suivante obtenue par intégration :

$$\forall t \in [t_b; t_e] \quad x^1r(t) + x^1l(t) = C_x \quad y^1r(t) - y^1l(t) = C_y$$

Nous déterminons dans un premier temps les constantes C_x et $\mathbf{C}_y$.

$$C_x = \sum_{t=t_b}^{t_e} \frac{x^0r(t) + x^0l(t)}{t_e + 1 - t_b}$$

$$C_y = \sum_{t=t_b}^{t_e} \frac{y^0r(t) - y^0l(t)}{t_e + 1 - t_b}$$

Il est ensuite possible de filtrer les coordonnées issues du suivi :

$$x^1r(t) = \frac{C_x}{2} + \frac{x^0r(t) - x^0l(t)}{2}$$

$$x^1l(t) = \frac{C_x}{2} + \frac{x^0l(t) - x^0r(t)}{2}$$

$$y^1r(t) = \frac{y^0r(t) + y^0l(t)}{2} + \frac{C_y}{2}$$

$$y^1l(t) = \frac{y^0l(t) + y^0r(t)}{2} - \frac{C_y}{2}$$

8.2.3.2 Utilisation du patron de mouvement

Une fois la symétrie utilisée, il est possible de prendre en compte le patron de mouvement qui est impliqué dans le signe.
Notre exemple de mouvement est balistique sans répétition. Nous choisissons d'approximer la tra-

jectoire par une parabole[4]. Nous supposons que les positions des mains P_b et P_e ont été estimées correctement au début et à la fin du signe.
Les calculs sont effectués dans le repère $(O, \vec{i'}, \vec{j'})$ dont le vecteur $\vec{i'}$ est colinéaire avec le vecteur $\overrightarrow{P_bP_e}$. Dans ce nouveau repère les nouvelles coordonnées de la main sont notées $(x'(t), y'(t))$.
Nous cherchons par regression linéaire la parabole d'équation $y'(t) = a.x'(t).x'(t) + b.x'(t) + c$ la plus proche de la trajectoire réelle au sens des moindres carrés, passant par les points P_b et P_e.

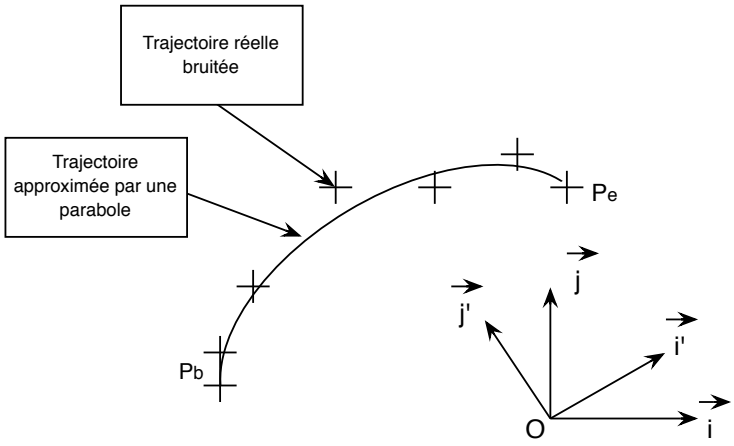


FIGURE 8.8 – Approximation de la trajectoire par une parabole

8.2.4 Etapes de création d'un photosigne

FIGURE 8.9 – Exemple de signe complexe : signe [ALSACE].

Les différentes méthodes de calcul que nous avons présentées dans la partie précédente ont été intégrées à un logiciel permettant d'aider à la création de photosignes. La version actuelle est déjà

[4]Pour l'instant, le modèle de flèche utilisé dans les photosignes de signes à mouvement balistique ne prend pas encore en compte l'asymétrie du profil de vitesse.

opérationnelle, mais ne permet de traiter que les signes impliquant des mouvements balistique, balistiques répétés, circulaires, circulaires répétés, en angle et en aller-retour. Une autre catégorie nommée “signes complexes” permet de représenter directement une trajectoire lissée de la main sous forme d’une flèche comme dans la figure 8.9. L’interface que nous avons réalisée permet de visualiser et de paramétrer le photosigne en cours de création. Cette interface est représentée dans la figure 8.10.

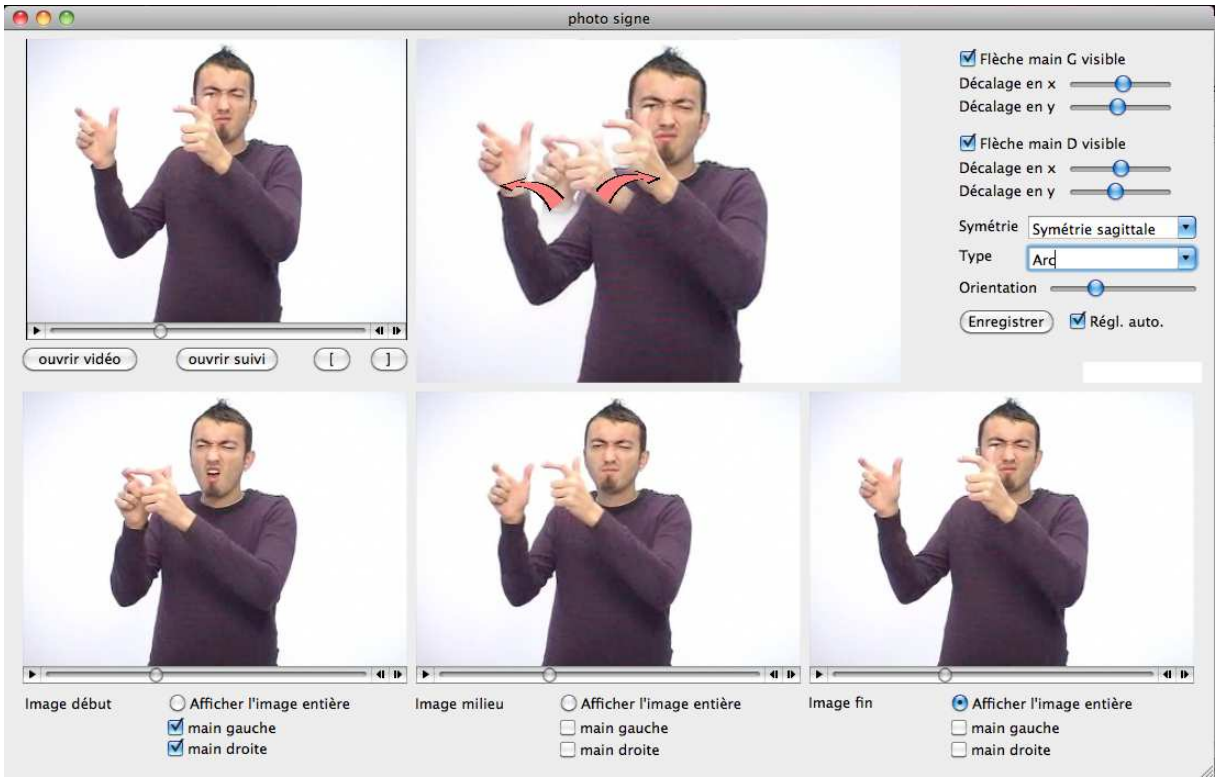


FIGURE 8.10 – Interface du logiciel Photosigne

La création d’un photosigne suppose plusieurs étapes d’interaction entre l’utilisateur et le logiciel :

1. Le suivi de la tête et des mains est réalisé grâce à l’algorithme décrit en §6. L’utilisateur doit juste spécifier la zone d’intérêt de la vidéo à suivre, la dimension de la tête, ainsi que l’échelle du signeur.

2. Il est ensuite nécessaire de sélectionner manuellement dans la vidéo, les images et les membres à faire apparaître en transparence dans le photosigne (image de début, image de fin, ou image en cours de signe).

3. L’utilisateur choisit ensuite le type de mouvement et de symétrie du signe (cette étape sera certainement automatisée prochainement en utilisant les algorithmes du §7.5).

4. La génération des flèches proprement dite est automatique, et ne nécessite donc pas d'intervention humaine. Elle est réalisée en traitant les données de suivi en fonction du type de signe.

5. Finalement, il est possible de déplacer les flèches dans l'image pour permettre une meilleure visibilité des mains du signeur.

L'application présentée est encore en cours de développement et n'a pas été évaluée de manière quantitative. Malgré tout, il est intéressant de mentionner un certain nombre de retours d'utilisateurs :

- Les résultats sont bons[5] pour les mouvements balistiques dont les emplacements de début et de fin de signe sont bien distincts.
- Lorsque les mains sont trop proches du visage, il existe des problèmes de transparence entre les différentes images sélectionnées par l'utilisateur (cf. figure 8.11).
- Les flèches ne traduisent pas les changements d'orientation de la main, qui sont pourtant une composante importante des signes.
- La forme actuelle des photosignes est peu adaptée à la représentation de signes impliquant uniquement des rotations du poignet, ou des changements de configuration. On se retrouve alors avec une superposition de deux configurations manuelles qui n'est guère satisfaisante visuellement (cf. figure 8.12).

Malgré tout, les résultats actuels apportent déjà un début de réponse à la problématique de représentation des signes qui permet une impression de trace 2D.

FIGURE 8.11 – Problème de transparence au niveau de la tête pour le signe [HIVER]

[5]Nous entendons par bon résultat, une images de signe esthétique qui permet aisément à celui qui la regarde de comprendre comment est réalisé le signe. Nous sommes toutefois conscient qu'une évaluation plus poussée nécessitera des critères plus objectifs.

FIGURE 8.12 – Problème de restitution d'un signe n'impliquant que des changements de configurations. Exemple du signe [VA-VA] indiquant le futur proche

8.3 Segmentation d'un énoncé en signes

Nous évoquions en introduction de ce chapitre, le besoin de pouvoir appliquer des modifications aux différents signes qui composent un énoncé, de pouvoir les sélectionner, les compter, les accorder … Ceci nécessite une décomposition de l'énoncé en un certain nombre de signes qui peuvent être traités indépendement les uns des autres. Deux cas de figure peuvent se présenter. Un énoncé peut être synthétisé à partir de signes placés les uns à côté des autres, au quel cas la segmentation est déjà réalisée. Il se peut aussi que l'énoncé ait été créé à l'aide de capture de mouvement ou de vidéo. Dans ce deuxième cas de figure, un traitement supplémentaire est nécessaire pour aboutir à une segmentation du signe.

Nous avons souligné dans le chapitre 3, le fait qu'il était très difficile de définir la notion de signe et, *a fortiori*, d'en déterminer le début et la fin. Nous commencerons à aborder la notion de pertinence de la segmentation d'une vidéo en signe. A l'aide de plusieurs expériences, nous montrerons qu'il est même possible d'approcher une segmentation correcte en utilisant uniquement le mouvement des signes.
Nous exposerons ensuite un algorithme de segmentation semi-automatique basé sur les patrons de mouvement et utilisant des contraintes entrées par l'utilisateur.
Pour finir, nous discuterons des étapes encore nécessaires pour aboutir à une segmentation totalement automatique d'un énoncé en signes.

8.3.1 Pertinence de la notion de segmentation

L'hypothèse que nous formulons est qu'il est possible de segmenter un énoncé en signes à partir du mouvement seul, en utilisant la phonologie des LS. Plusieurs questions en découlent :

- La notion de segmentation d'un énoncé en signes a-t-elle un sens ?

- Peut-on segmenter un énoncé à partir de la seule information de mouvement ?

- Peut-on segmenter un énoncé sans le comprendre ?

L'expérience qui permet de répondre à ces questions n'est pas directement en lien avec le traitement d'image, mais nous permet de valider des hypothèses que nous utiliserons dans le cadre de la segmentation semi-automatique.

8.3.1.1 Qu'est-ce qu'une bonne segmentation ?

Il est très difficile de définir à quel instant commence ou finit un signe car la segmentation dépend du critère utilisé (point de rebroussement du mouvement, stabilisation de la configuration, labialisation, expression du visage). Nous essayons donc de définir ce qu'est une segmentation acceptable. Il s'agit ici d'une proposition de critères que nous soumettons à la communauté d'informaticiens et de linguistes. Les définitions qui suivent seront certainement amenées à évoluer en fonction des connaissances futures qui seront développées sur les LS.
Un signe est défini par un ensemble de traits (orientations, configurations, mouvements, emplacements ...) et de relations qui existent entre les différents articulateurs (symétrie, conservation d'orthogonalité ...). Nous appelons **noyau** du signe, le plus petit intervalle temporel de la réalisation du signe dans lequel tous les traits du signe sont présents. Nous appelons **support**, le plus grand intervalle durant lequel les mouvements de tous les articulateurs du signeur sont compatibles avec les traits du signe[6]. Les mouvements présents dans le noyau sont en général précédés d'un mouvement de préparation et suivis d'une tenue ou d'une phase de retrait qui peuvent faire partie du support du signe. Deux signes peuvent partager une partie de leur support. Par contre, il est extrêmement rare qu'il y ait intersection entre le noyau d'un signe et les supports des signes adjacents[7]. Les relations temporelles entre noyaux et supports sont représentés sur la figure 8.13.

[6]Nous utilisons volontairement une terminologie liée aux ensembles flous qui permettent de quantifier le degré d'appartenance d'un élément à un ensemble.

[7]Les seules exceptions que nous avons trouvées concernent les signes composés dans lesquels les deux signes sont partiellement imbriqués comme dans le signe [PENSER DIFFÉREMMENT].

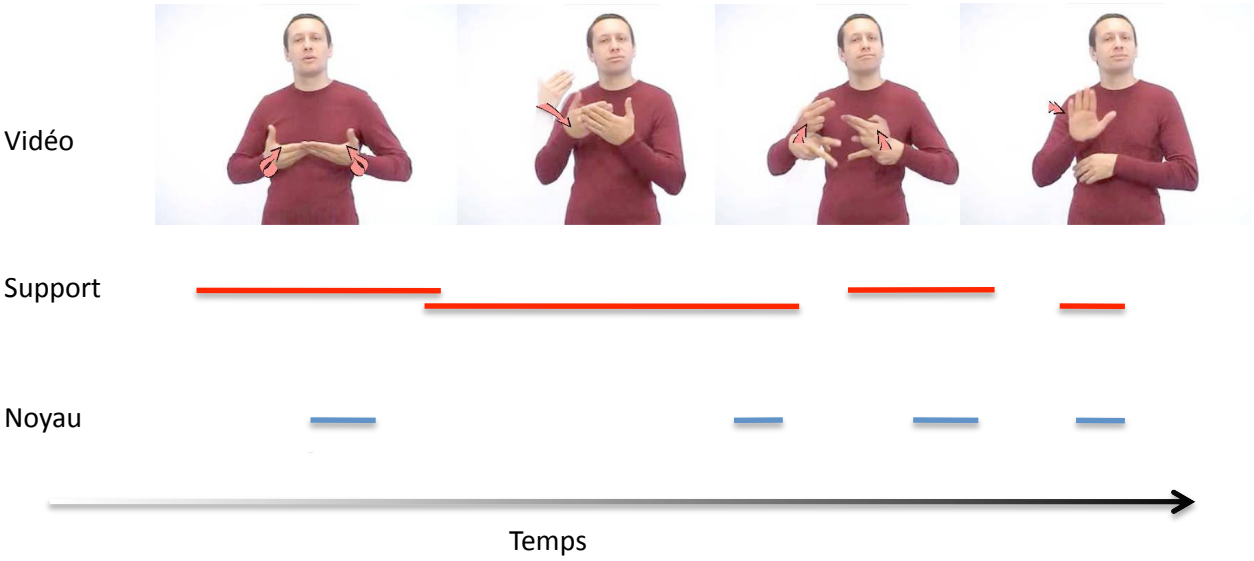


FIGURE 8.13 – Relations temporelles entre les noyaux et les supports des signes.

8.3.1.2 Protocole

Nous avons choisi une vidéo de 42 secondes issue d'une traduction de brève d'actualité par l'entreprise Websourd pour effectuer les expériences de segmentation. Dans un premier temps, un suivi de la tête et des mains est effectué. Les coordonnées successives sont utilisées pour créer une vidéo dans laquelle la tête et les mains du signeur sont figurées par des cercles colorés.
On demande à chaque expérimentateur de segmenter la vidéo, où seul le mouvement est visible, puis de segmenter la vidéo d'origine.
Nous utilisons le logiciel SLAnnotation (cf. figure 8.14) pour mener à bien la segmentation. Ce logiciel permet de délimiter des segments temporels de la vidéo, et d'y adjoindre des annotations sous forme textuelle ou vidéo. S'il pense avoir reconnu le signe qu'il est en train de segmenter, l'expérimentateur est invité à effectuer ce signe devant une caméra et la vidéo est ajoutée comme annotation du segment. Ces annotations visuelles permettent d'identifier le type de mouvement que le signeur pense avoir reconnu. Nous nous assurons par la même occasion que le signeur n'utilise pas une compréhension, même partielle de l'énoncé, pour aboutir à une segmentation.
Quatre annotateurs ont participé à l'expérience :

- Une collaboratrice sourde signante (A) habituée à effectuer des tâches d'annotation de vidéos en LS,

- Une personne sourde signante (B) en formation pour enseigner la LSF,

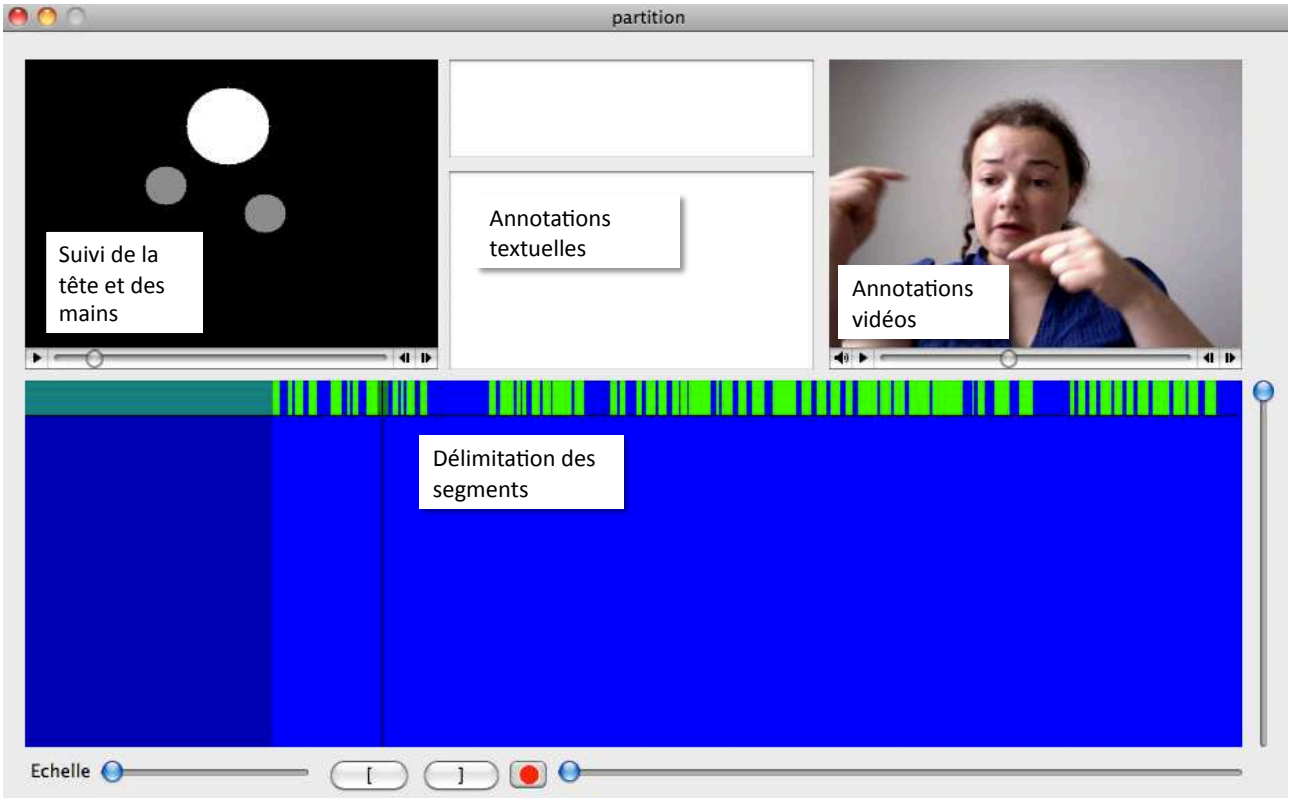


FIGURE 8.14 – Logiciel SLAnnotation utilisé pour segmenter un énoncé à partir des mouvements et enregistrer des annotations vidéos.

– Une collaboratrice entendante (C), débutante en langue des signes, habituée à effectuer des annotations de vidéos en LS,

– Une personne entendante (D) n'ayant été en contact que rarement avec des sourd, ne connaissant pas la LSF.

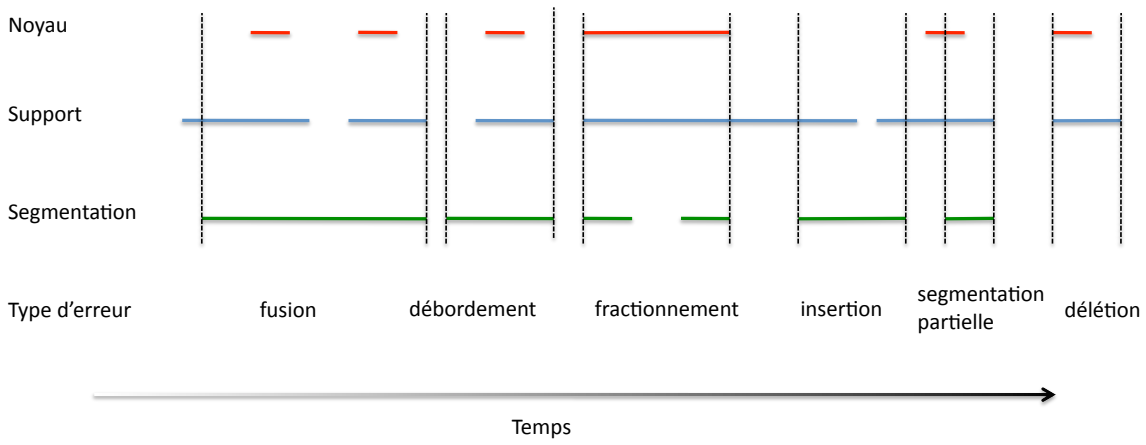


FIGURE 8.15 – Catégories d'erreurs de segmentation

Pour définir les noyaux et les supports des signes, nous choisissons de ne prendre en compte que les traits du signes relatifs au mouvement et à l'emplacement des mains du signeur. Les erreurs de segmentation sont classées de la manière suivante (cf. figure 8.15) :

- Les **délétions** correspondent aux signes dont le noyau est absent dans le segment.
- Les **insertions** sont des segments qui ne contiennent aucun noyau.
- Les **segmentations partielles** sont des segments qui ne contiennent qu'une partie du noyau du signe (par exemple, il manque un contact). Nous nommerons **fragmentation** le fait qu'un noyau soit partiellement présent dans deux segments. La fragmentation est un cas particulier de segmentation partielle.
- Les **débordements** désignent les segments qui vont au delà du support du signe. Parmi eux, les **fusions** désignent les segments qui contiennent plusieurs noyaux de signes.

Il est fréquent que les erreurs de segmentation appartiennent à la fois aux catégories *segmentation partielle* et *débordement*.

8.3.1.3 Résultats

Nous observons une concordance des deux segmentations des expérimentatrices (A) et (B), obtenues à partir de la vidéo complète. Les segmentations en signes concordent pour 55 des 56 signes, à l'exception d'un nom de ville qu'une annotatrice a segmenté comme un seul signe et que l'autre a divisé en deux signes.

Nous notons également une concordance entre la segmentation fournie par la personne entendante non signante (D) avec celle proposée par l'annotatrice sourde signante (B) à partir de la vidéo complète. Parmi les 56 signes, seuls les signes [ENVIRON], [NEUF], [CINQ], [MAINTENANT], [PREMIEREMENT], [DEUXIEMEMENT], [TROISIEMEMENT], [QUATRIEMEMENT] et une épellation ont été reconnus. Nous observons deux délétions, une fragmentation, un débordement ainsi que trois fusions. Les 48 signes restants ont été correctement segmentés. En supprimant les signes dont le sens avait été deviné, 39 signes (soit 70 % des signes) ont été correctement segmentés. Interrogée sur sa méthode de segmentation, l'expérimentatrice (D) a décrit la stratégie suivante :

- Un signe correspond en général à une configuration manuelle stable,
- La configuration manuelle peut changer à l'intérieur d'un signe, mais le changement de configuration doit être cohérent avec le mouvement effectué par le signeur,
- Un signe implique soit une main, soit les deux mains du signeur,
- L'expression du visage du signeur est maintenue durant l'ensemble de la durée du signe.

Les résultats de la segmentation à partir du mouvement seul sont beaucoup plus variables d'un expérimentateur à l'autre. Nous évaluons la justesse de la segmentation en termes de précision et de rappel en utilisant les deux formules suivantes :

$$précision = \frac{nombre\ de\ signes\ correctement\ segmentés}{nombre\ de\ segments\ proposés}$$

$$rappel = \frac{nombre\ de\ signes\ correctement\ segmentés}{nombre\ de\ signes\ à\ segmenter}$$

L'expérimentatrice (A) aboutit à une précision de segmentation 51% et un rappel de 46%. Les hypothèses de signes fournies à partir des mouvements ont montré que la signeuse se basait bien sur la dynamique du mouvement pour aboutir à une segmentation (les mouvements proposés étaient de la même catégorie que les mouvements réels). Trois signes ont été devinés à partir du mouvement sans pour autant que le sens de l'article puisse être inféré.
L'expérimentatrice (B) obtient une précision de 36% ainsi qu'un rappel de 27%. Les hypothèses de signes filmées par la webcam montrent que les signes proposés ont le même emplacement que les signes réels, sans être pour autant de la même catégorie de mouvement. On observe en général une sous-segmentation de l'énoncé qui se manifeste sous forme de fusions ou de débordements. Aucun des signes n'a été deviné.
L'expérimentatrice (C) propose une segmentation d'une précision de 28% et d'un rappel de 41%. Les hypothèses de signes prouvent que la signeuse se base bien sur la dynamique du signe puisque les signes proposés sont de la même catégorie que les signes de la vidéo à segmenter. Aucun des

signes n'a été deviné. Globalement, on observe une sur-segmentation de l'énoncé qui se manifeste sous la forme de nombreuses insertions de signe (souvent il s'agit d'une transition segmentée comme un signe).
L'expérimentatrice (D) obtient une précision de 18% et un rappel de 11%. Nous avons demandé à l'expérimentatrice de proposer des mouvements stéréotypiques (se brosser les dents, jouer du tambour, ouvrir une porte) qui pourraient correspondre aux gestes segmentés. Les mouvements proposés avaient le même emplacement que les signes réels, mais n'appartenaient pas à la même catégorie de mouvement. Les critères utilisés pour la segmentation étaient, de l'avis de l'expérimentatrice, le changement d'emplacement et la présence de pauses entre les signes. On observe globalement une sous-segmentation qui aboutit à un grand nombre de fusions et de débordements.

Il ressort des expériences que les signes n'impliquant pas de mouvement ne sont que très rarement segmentés correctement.

8.3.1.4 Conclusions

Avant de conclure à partir de ces observations, soulignons le fait que la même expérience devrait être répétée avec davantage d'expérimentateurs en utilisant des vidéos plus diversifiées et plus longues. Les données de suivi obtenues à l'aide de notre algorithme, comportaient quelques erreurs qui ont parfois induites les segmentateurs en erreur. Il serait nécessaire de réitérer l'expérience avec des données de capture de mouvements.
Les résultats préliminaires que nous observons permettent toutefois de valider déjà plusieurs hypothèses :

- La concordance entre les segmentations d'un énoncé de plusieurs sourds signeurs prouve la pertinence de la notion de segmentation d'un énoncé lorsqu'il contient des signes standards. Les résultats d'autres expériences de segmentation que nous avons menées sur des vidéos de narration montrent toutefois une plus grande variabilité des résultats de segmentation pour les unités gestuelles de grande iconicité.

- Le fait qu'une personne ne pratiquant pas la LSF soit en mesure de segmenter correctement l'énoncé prouve qu'il est possible de segmenter un signe en se basant uniquement sur sa structure, sans faire intervenir sa signification. Les règles de segmentation formulées par l'expérimentatrice non-signante sont d'ailleurs relativement proches des règles phonologiques proposées par les linguistes.

- Les segmentations basées uniquement sur le mouvement montrent qu'il est possible de segmenter correctement certains signes à partir d'un simple suivi de la tête et des mains. L'utilisation de la catégorie du signe (balistique, circluaire ...), semble être plus pertinente que celle de

variation d'emplacements ou de pause entre les signes. L'utilisation du mouvement seul peut mener à de nombreuses confusions entre signes et transitions.

8.3.2 Segmentation semi-automatique

Les expériences que nous avons menées vont dans le sens d'une segmentation possible en utilisant uniquement l'information de mouvement. Elles mettent en même temps en lumière la difficulté de distinguer les signes des transitions.
Nous cherchons ici à automatiser partiellement la segmentation en fournissant à l'algorithme de segmentation des informations supplémentaires. Le but est de fournir à d'autres outils, des segments de vidéos utilisables pour effectuer des opérations de traitement de signe comme l'indexation ou la production de photosignes. Nous présentons une méthode permettant de résoudre ce problème. L'évaluation que nous proposons est basée sur la segmentation assistée d'une narration comportant des signes standards et des unités gestuelles iconiques. Nous présentons quantitativement et qualitativement les résultats de cette segmentation, puis nous proposons plusieurs améliorations qui permettraient de fournir des segmentations plus précises et de se passer des informations fournies par l'utilisateur.

8.3.2.1 Méthode de segmentation assistée

La méthode de segmentation que nous proposons se déroule en quatre étapes :

- L'utilisateur visualise la vidéo à vitesse réelle et presse une touche à chaque fois qu'il identifie un signe. Nous nommons "amorces" ces évènements.

- On effectue un suivi des mains du signeur dans la vidéo de manière à connaître à chaque instant leur position 2D dans l'image.

- L'algorithme propose une segmentation de la vidéo en utilisant les données du suivi, ainsi que les patrons de mouvements.

- L'utilisateur valide et corrige si besoin la segmentation fournie par l'utilisateur.

Nous avons défini au chapitre 7 une méthode permettant de définir une mesure d'appartenance d'un segment temporel seg_{tl} commençant à l'instant t et de durée l, à une catégorie de mouvements. Il est possible de définir une mesure de confiance p_{tls} indiquant l'appartenance de ce segment à la classe

signe comme le maximum de toutes les mesures d'appartenance du segment aux catégories de mouvements[8]. De la même manière, nous créons une mesure de confiance p_{tlt} indiquant l'appartenance du segment à la catégorie transition.
Les opérateurs pris en compte pour déterminer l'appartenance du segment à la classe signe portent sur la détection de symétrie ou de main statique, la détection de répétitions dans les signes, la dynamique du signe (régularité de la vitesse ou accélération et décélération brusque). Pour des raisons de concision, nous ne décrivons pas le principe de fonctionnement des différents opérateurs que le lecteur pourra trouver dans [LAGD08].
Un énoncé est modélisé comme étant une alternance de signes et de transitions. Il s'agit de maximiser la vraisemblance V_s de la segmentation sous la contrainte que chaque signe contienne une et une seule amorce et que chaque transition n'en contienne aucune. La vraisemblance est calculée à l'aide de la formule suivante dans laquelle i représente le numéro du segment. p_{tli} et l_i correspondent respectivement à la mesure de confiance des signes ou des transitions et à la durée de ces segments :

$$V_s = \sum_i ln(p_{tli}).l_i$$

Cette optimisation sous contrainte est réalisée en utilisant un algorithme de programmation dynamique.

8.3.2.2 Evaluation

Nous nous plaçons dans le cadre d'une segmentation réalisée dans un but d'être réutilisée pour créer des photosignes ou constituer une succession de vidéos autonomes (utilisées par exemple dans des tables des matières). Il faut donc que chaque segment contienne l'intégralité des traits du signe, sans comporter de traits des signes voisins. Cela revient à dire que la segmentation de chaque signe doit être incluse dans le support du signe et inclure son noyau. Les noyaux et les supports des signes sont cette fois définis en prenant en compte à la fois les mouvements, les configurations, les emplacements et les positions des mains.
Une segmentation sera par exemple déclarée comme non satisfaisante dans les cas suivant :

- Il n'y a pas assez d'images dans le segment pour pouvoir déterminer le mouvement du signe.
- Les répétitions ou les allers-retours ne sont pas visibles dans le segment.
- Le nombre de contacts n'est pas respécté.

[8]Le développement de l'algorithme de segmentation que nous présentons est antérieur aux modèles de signes que nous avons présenté, les patrons géométriques et dynamiques utilisés par l'algorithme sont donc légèrement différent de ceux présentés au chapitre 7. En particulier, ils ne prennent pas en compte les phases de préparation et de retrait. Les opérateurs de détection de symétrie sont également légèrement différents.

– Une partie d'un signe adjacent (mouvement ou configuration) est visible dans le segment.

Tous les segments temporels appartenant à ces cas de figure seront classés dans la catégorie *segmentation partielle*, les autres seront considérés comme satisfaisants. Cela signifie que plusieurs segmentations différentes seront possibles pour un même signe. Dans la plupart des cas, cette différence de segmentation ne se manifestera que par une variation de l'amplitude du signe.

Notre algorithme de segmentation est validé sur une séquence de deux minutes produite dans le cadre de la réalisation du corpus LS Colin [Cux02]. Il s'agit d'une séquence dans laquelle le locuteur décrit les attentats du 11 septembre 2001. Cette séquence a été segmentée manuellement par une collaboratrice sourde signante et contient 203 signes.
Nous utiliserons alternativement des mesures des positions 2D des mains dans la vidéo issues de notre logiciel de suivi des mains, et des mesures effectuées manuellement. Ceci nous permettra d'évaluer l'influence de la précision des position des mains sur la segmentation.
Nous utiliserons également un fichier d'amorces saisi manuellement lors du visionnage de la vidéo ainsi qu'un fichier où chaque amorce correspond à un instant quelconque de l'intervalle du signe correspondant segmenté manuellement.
A titre de comparaison, nous présenterons les résultats d'une segmentation automatique effectuée par notre logiciel de segmentation sans prendre en compte le fichier d'amorces. Dans chaque cas de figure, on distinguera les signes segmentés correctement (bon), les signes segmentés partiellement (partiel) et les signes de la segmentation manuelle qui ne sont identifiables dans aucun des segments temporels issus de la segmentation assistée (del). Les résultats sont présentés dans le tableau 8.16.

fichier d'amorces	sans			manuel			manuel		
positions des mains	manuel			manuel			auto		
signes segmentés	bon	partiel	del.	bon	partiel	del.	bon	partiel	del.
	48	121	34	82	97	24	87	80	36

fichier d'amorces	auto			auto		
positions des mains	manuel			auto		
signes segmentés	bon	partiel	del.	bon	partiel	del.
	76	116	11	75	120	8

FIGURE 8.16 – Résultats expérimentaux issus du logiciel de segmentation assistée

Plusieurs remarques intéressantes peuvent être formulées sur ces résultats :

– L'erreur sur la position 2D des mains due au logiciel de suivi des mains et du visage n'a que peu d'influence sur la précision de la segmentation. L'écart de signes segmentés correctement est dans tous les cas inférieur à 5% par rapport au nombre de signes total. Ceci confirme la robustesse des opérateurs de traitement du mouvement que nous utilisons.

– L'utilisation d'un pointage manuel des signes, même approximatif, permet dans un temps relativement restreint de doubler le taux de segmentation correcte des signes.

Une observation plus troublante au premier abord est que les résultats obtenus avec un fichier d'amorce automatique sont moins bons que ceux issus du fichier d'amorce manuel. Ceci a deux origines :

– Une part significative des signes non repérés lors de la création du fichier d'amorce manuel étaient en fait des hésitations ou des signes extrêmement courts. A titre indicatif, ces signes non repérés avaient une durée moyenne de 6 images, contre une durée moyenne de 17 images pour l'ensemble des signes. Ces signes sont passés de la catégorie *délétion* à la catégorie *partiellement segmenté* sans pour autant être délimités de manière satisfaisante.

– On peut aussi émettre l'hypothèse que la personne qui a saisi les amorce, les a situées préférentiellement dans les noyaux des signes alors que les amorces générées automatiquement étaient placées à un instant quelconque dans leur support.

8.3.3 Pistes d'amélioration

Les résultats obtenus sont encourageants compte tenu du peu d'informations prises en compte pour réaliser cette segmentation temporelle. Plusieurs améliorations seront nécessaires pour permettre une exploitation de ce procédé dans des logiciels de traitement de signes.
Etant donné la méthode de segmentation mise en œuvre, il nous est possible de qualifier le résultat de la segmentation et d'indiquer à l'opérateur humain la confiance qu'il peut avoir dans la segmentation de chaque signe. Cela permettrait d'accélérer la vérification de la segmentation en évitant de demander de valider les signes ayant un indice de confiance élevé.

Actuellement, près de 20% des signes présents dans l'énoncé original ne sont identifiables dans aucun des segments temporels produits par la segmentation assistée. Cela entraine une perte de temps considérable lors de la vérification car il faut alors visualiser l'ensemble de la vidéo pour retrouver ces signes oubliés. Il serait donc nécessaire de trouver d'autres méthode pour détecter ces signes, au moins partiellement, afin que l'étape de vérification se limite à une visualisation des intervalles temporels correspondant aux signes segmentés. Cela nécessiterait de prendre en compte d'autres indices que le mouvement. Le plus utilisé pour la segmentation par un opérateur humain est certinement la configuration manuelle mais elle est extrêmement difficile à extraire automatiquement. D'autres indices comme les mouvements de la tête, la direction du regard ou les clignements d'yeux pourraient en revanche être plus simples à extraire et à utiliser pour améliorer les résultats de segmentation.

8.4 Bilan

Nous venons d'aborder dans ce chapitre la notion d'outils de traitement de signes, puis nous avons approfondi les notions de création automatique d'images de signe et de segmentation des énoncés.
Pour clore ce chapitre et cette partie de la thèse sur nos contributions, nous reprenons les différentes fonctionnalités que peut offrir un logiciel de traitement de texte, et nous montrons en quoi nos travaux ont permis de proposer des méthodes pour effectuer des traitements similaires dans des énoncés en LSF. Nous faisons naturellement référence aux différents modèles et aux différentes méthodes présentées dans cette thèse, mais nous évoquons également d'autres outils, réalisés durant ces trois ans qui sont davantage décrits dans la publication [LADD$^+$re].
Les solutions décrites dans le cadre de cette thèse peuvent avoir les fonctions suivantes :

Segmenter un énoncé en signes La modification d'un énoncé est d'autant plus simple qu'on peut le segmenter en signes et donc, sélectionner les signes plus facilement.

Modifier un signe La caractérisation d'un signe permet de connaître ses paramètres de variabilité, et de pouvoir proposer à l'utilisateur un certain nombre d'opérations pour modifier le signe.

Rechercher un signe L'architecture que nous avons proposée permet de localiser plus facilement un signe dans un document.

Annoter un document Nous avons créé dans le cadre du projet SESCA, deux logiciels (AVV et SLAnnotation) permettant de commenter en LS, un document sous forme vidéo.

Modifier le style d'un signe nous avons identifié un certain nombre de paramètres qui peuvent varier lors de la réalisation d'un signe, en montrant également les liens qui existent entre eux. L'approfondissement de cette analyse permettra peut-être, à moyen terme, de dégager certains paramètres qui ont trait au style du signeur. La modification systématique des profils de vitesse, de la rapidité d'exécution des signes ou de l'amplitude pourrait peut-être changer le style d'expression sans affecter le sens de l'énoncé produit.

Imprimer un signe Nous proposons une solution permettant de créer des images à partir de signes et de les rendre imprimables.

Permettre la dictée automatique de signe Nous ne résolvons que partiellement le problème, car nous ne suivons que les positions des mains, des coudes, de la tête et du buste. Par contre, la robustesse du système de suivi permet déjà des applications opérationnelles.

Les fonctionnalités que nous venons d'énumérer ainsi que les autres possibilités offertes par les outils d'infographie constituent des briques qui permettront certainement de construire, à terme, de

véritables outils destinés à la production et à la mise en forme d'énoncés en LSF, respectant toutes les particularités de cette langue. Pour l'instant toutefois, il ne s'agit que de modules distincts qui n'ont pas encore été regroupés dans un même logiciel. L'intégration des différentes fonctionnalités au coeur d'une même application nécessitera une unification des modèles de représentation des signes qui constitue un défi d'avenir.

Quatrième partie

Conclusion et perspectives

CHAPITRE 9

PERSPECTIVES

La route est encore longue avant d'aboutir à un traitement automatique des LS aussi perfectionné que celui qu'on atteint actuellement avec les langues vocales. Voici un certain nombre de perspectives ouvertes par cette thèse. Certaines relèvent de notre domaine de traiteur d'image, tandis que d'autres concernent plutôt la synthèse par signeur virtuel.

9.1 Traitement automatique de vidéos en LS

9.1.1 Aller plus loin dans le suivi

Nous avons présenté dans le chapitre 6, un algorithme permettant d'effectuer le suivi de différents articulateurs dans la vidéo. L'évaluation a été l'occasion de mettre en lumière plusieurs problèmes récurrents :

- Une difficulté à localiser la main lors d'une occulation totale de la tête ou lors d'une sortie de cadre prolongée des mains du signeur (problème d'autant plus saillant lorsque ces évènements sont de longue durée),
- Une perte de précision sur la localisation de la main lorsque le signeur porte des manches courtes et que la main du signeur occulte sa tête,
- Une désambigüisation des mains encore imparfaite, malgré des améliorations significatives résultant de la prise en compte d'indices linguistiques.

Ces problèmes montrent les limites d'une approche basée uniquement sur la détection de couleur de peau et de fond. Pour améliorer le suivi, il sera certainement nécessaire de prendre en compte d'autres informations.
La prise en compte de la **texture** permettrait de suivre la main du signeur de manière plus précise, même en présence d'occultations de la tête.
La prise en compte de la **forme des mains** serait certainement aussi un moyen efficace pour désambigüiser les deux mains. Les résultats présentés par [BEZre], qui compare les formes de la mains à des configurations manuelles de référence reposent sur cette approche et sont tout à fait encourageants.
De toute façon, l'utilisation de la forme et de la texture deviendra incontournable, dès lors qu'on s'intéressera à la caractérisation ou à la classification automatique d'autres paramètres rentrant en compte dans une production signée, comme l'identification des configurations manuelles, l'estimation des rotations de la tête et la caractérisation de l'expression du visage comme le propose [PHD10].

9.1.2 Des modèles de signe plus complets

L'une des spécificités de cette thèse d'analyse est le développement de modèles permettant de caractériser les mouvements des signes. De nombreuses améliorations sont encore à apporter à ces modèles :

- Nous avons traité uniquement les cas des mouvements balistiques avec, et sans répétition. Il serait nécessaire de mener ce travail de modélisation paramétrique pour l'analyse avec les autres types de mouvements les plus répandus (croix, cercle, angle ...).
- Il existe des différences sur le plan de la dynamique, entre les transitions et les mouvements balistiques. Nous avons réussi à mettre en évidence l'asymétrie des profils de vitesse. Il serait intéressant de pousser plus loin l'investigation, pour voir si d'autres paramètres permettent d'aboutir à une distinction.
- Nos modèles de mouvements ont uniquement porté sur la trajectoire et la dynamique des mains du signeur. Il est vraisemblable que les mouvements des coudes soint aussi coordonnés avec ceux des mains. La création d'un modèle de mouvement pour l'analyse incluant à la fois la dynamique des coudes et des mains pourrait comporter deux avantages. L'utilisation du modèle permettrait de rendre le suivi plus robuste aux occultations, en prenant en compte la redondance d'informations entre les déplacement des coudes et ceux des mains. Le fait que le déplacement des coudes précède légèrement celui des mains permettrait de détecter plus rapidement certains signes.

Au delà du modèle de mouvement, il pourrait être intéressant de développer des modélisations génériques de signes orientées vers l'analyse. Nous savons d'après des travaux de Battison [Bat74] que de nombreuses contraintes comme la symétrie pèsent à la fois sur le mouvement, sur les orientations et sur les configurations manuelles. On pourrait donc imaginer des modèles de signes pour l'analyse, qui prennent en compte ces relations d'interdépendance.

9.1.3 Vers un traitement descendant plus généralisé

Notre travail a été l'occasion de montrer qu'un traitement descendant de la vidéo était possible. Nous exposons trois mécanismes qui permettent d'aboutir à une correction des résultats du suivi, en utilisant la phonologie de la LSF :

- La désambigüisation des mains droite et gauche du signeur en utilisant la notion de main dominante,
- La correction de la profondeur en utilisant la corrélation entre les profondeurs des mains droite et gauche,

– L'idéalisation des trajectoires de suivi en utilisant les patrons de mouvement.

Il serait certainement possible d'aller plus loin dans cette approche descendante. Au lieu d'utiliser une corrélation moyenne entre les profondeurs des deux mains, nous pourrions adapter la méthode de correction à partir des relations de symétrie entre main droite et gauche, estimées dans le plan image.
D'autre part, le même principe qui a permis d'aboutir à une création de trajectoire idéale à partir de trajectoires réelles bruitées, pourrait certainement être aussi utilisé pour débruiter des données de suivi.

Au delà du traitement du mouvement, l'approche d'analyse de la vidéo, guidée par un modèle de LS pourrait également s'appliquer à d'autres paramètres. Si on part du principe qu'il existe un lien entre la structure d'un mouvement et les contraintes sur les configurations, il sera peut-être possible d'émettre des hypothèses sur le nombre maximum de configurations que contiendra un signe, les images qu'il faut traiter pour reconnaître la configuration ou les enchaînements possibles de configurations. Cette approche descendante est d'autant plus nécessaire que l'opération de reconnaissance de configurations est très chronophage et qu'elle est rendu complexe par le flou des images, les occultations et le grand nombre de degrés de mobilité de la main.

Les expériences de segmentation manuelle ont permis d'aborder avec les expérimentateurs, les différents critères utilisés pour déterminer le début et la fin des signes. Il s'avère que les mouvements de la tête, le changement d'expression du visage, la direction du regard et les clignements d'yeux étaient des indices utilisés comme marqueurs de fin de signe. Peut-être serait il possible d'utiliser ces paramètres non-manuels dès l'étape d'analyse bas niveau ?

9.1.4 Pour aller plus haut

Se rapprocher de l'avenir, c'est penser le traitement bas niveau que nous venons d'exposer, intégré dans un système de traitement automatique prenant en compte la totalité de l'énoncé, et permettant une modélisation de plus haut niveau. Comment passer d'un niveau d'analyse lexical à une analyse de la syntaxe, de la sémantique ou de la pragmatique d'un énoncé ? Nous donnons, dans les lignes qui suivent, plusieurs pistes pour tirer parti de l'analyse paramétrique que nous proposons.

9.1.4.1 Aide à la modélisation de l'espace de signation

La prise en compte de la spatialisation d'un énoncé est indispensable dès lors qu'on souhaite faire une analyse de la structure d'un énoncé en LS. De ce point de vue, les modèles de mouvements que nous présentons dans notre travail présentent plusieurs avantages :

- L'estimation systématique de l'emplacement du signe dans l'ensemble des paramètres caractérisant le mouvement permet de connaître précisément l'emplacement d'un concept dans l'espace de signation (ES) dans le cas où le signe le désignant est spatialisé par les mains.

- L'estimation de l'emplacement et de l'orientation des signes peut permettre d'interpréter les verbes en inférant l'émetteur et le destinateur d'une action.

- L'estimation de l'amplitude des mouvements, des positions relatives des mains du signeur, de l'orientation du mouvement, de sa courbure peut permettre de connaître plus précisément l'étendue d'un objet ou d'une zone dans l'ES, dans le cadre d'une énonciation en transfert situationnel.

Nous savons dès à présent utiliser ces informations dans le cadre d'une reconstruction assistée de l'ES. Nous avons eu l'occasion de réaliser un prototype de logiciel [LADD$^+$re] dans lequel l'utilisateur effectue une analyse de la structure grammaticale de l'énoncé, en indiquant à quel instant les concepts ou les actions sont placés dans l'espace. Le logiciel permet ensuite d'utiliser les données de suivi pour placer les concepts automatiquement dans l'espace de signation en utilisant la position et le mouvement des mains. Le résultat peut être visualisé sous la forme d'une vidéo sur laquelle la structure schématique de l'énoncé apparaît en transparence (cf. figure 9.1).

FIGURE 9.1 – Figuration de la structure diagrammaticale de l'énoncé par le biais de réalité augmentée.

Il reste donc maintenant à effectuer automatiquement le travail effectué par l'utilisateur. Cette tâche est colossale et demandera de résoudre plusieurs problèmes :

- Reconnaître les signes standards malgré leur variablité,

- Déterminer automatiquement la localisation des concepts par d'autres méthodes que la position des mains (orientation du buste, orientation de la tête, orientation du regard, utilisation d'un pointage),

- Segmenter correctement les signes non-standards et les catégoriser (action, personne, lieu ...),
- Détecter les prises de rôle et arriver à distinguer les personnages transférés d'après l'orientation du buste, l'expression du visage et la prosodie,
- Modéliser l'oubli (tous les objets situés dans l'espace de signation ne demeurent pas *ad vitam aeternam* s'ils ne sont pas réactivés),
- Tenir compte de la projection simultanée de plusieurs espaces conceptuels dans l'espace de signation (axe des temps, relations spatiales, relations hiérarchiques, relations causales ...),
- Faire automatiquement l'association entre les différentes entités localisées dans l'espace de signation (par exemple : le personnage P qui vient d'être placé dans l'espace de signation est joué dans le cadre d'un transfert personnel, l'objet A se trouve dans le lieu B ...),

9.1.4.2 Interprétation sémantique des paramètres

Notre approche paramétrique de description des signes permet non seulement d'estimer la distance entre deux signes, mais également d'estimer les paramètres qui varient entre deux exécutions d'un même signe. Comme nous l'avons dit précédemment, cette variation peut être interprétée au niveau syntaxique. Elle peut également apporter des nuances sur le sens du signe. Comme nous le soulignions au chapitre 3, il est possible en faisant varier l'écartement entre les mains, l'orientation du mouvement ou l'amplitude du signe [IMMEUBLE], de signifier qu'un immeuble est large, penché ou haut. Il serait donc imaginable de réaliser à long terme, une traduction d'énoncés en LS, tenant compte de ces nuances.
Là aussi, même si le principe est séduisant, de nombreux obstacles restent à surmonter :

- Arriver à distinguer, d'une part, les variations dans l'exécution du signe qui relèvent du style du signeur (ou de la personne qu'il transfère), de la prosodie, de la variabilité naturelle du signe, et d'autre part, les variations qui apportent réellement une nuance de sens au signe,
- Savoir comment interpréter les variations en terme sémantique (Un ralentissement peut signifier que l'action est longue dans certains signes d'action, et peut également souligner la taille importante d'un objet),
- Arriver à combiner les paramètres manuels aux paramètres non-manuels pour en proposer une interprétation correcte (la difficulté sera très importante pour l'interprétation des expressions du visage).

Il est encore loin, le temps où un signeur enregistrera un petit énoncé en LSF en utilisant des structures de grande iconicité, et où l'ordinateur annoncera dans une synthèse vocale impeccable : "L'homme avançait prudemment en regardant loin devant lui".

9.2 Couplage entre différents modèles pour l'analyse et la synthèse

Nous proposons dans cette thèse, un modèle paramétrique pour l'analyse de signes. Certains des paramètres que nous mentionnons comme la différence des amplitudes entre les différents mouvements impliqués dans une répétition ne sont pas présents dans la plupart des modèles linguistiques décrivant les signes. De plus, peu de modèles vont jusqu'à s'intéresser aux phases de préparation et de retraits des signes. Nous aurions tort de conclure prématurément qu'il s'agit d'une supériorité de notre modèle.
Il est en fait possible de distinguer deux catégories de modèle :

- Les premiers cherchent uniquement à représenter l'intention du signeur. Ces modèles peuvent naturellement tenir compte de la variabilité du signe en terme de flexion spatiale et temporelle, mais ils décrivent une réalisation idéale des signes non affectée par les phénomènes de coarticulation. Dans ces modèles de type A, seul le noyau du signe est décrit.

- Les seconds essaient de modéliser l'ensemble du signe tel qu'il est produit. Ainsi, l'ensemble du support du signe peut être décrit. Nous situons nos modèles de mouvements dans cette catégorie de modèles de type B.

Notons que les modèles de type A ou B ne sont pas forcément liés à une application d'analyse ou de synthèse de LSF et que les deux types de modèles seront certainement impliqués dans plusieurs tâches de reconnaissance ou de génération automatique d'énoncés en LSF. Nous proposons dans la section qui suit différents couplages qui peuvent exister entre les deux modèles.

9.2.1 Couplage des modèles pour la reconnaissance de signes

Dans le cas de reconnaissance, nous disposons de signes en contexte et nous souhaitons isoler chacun des signes afin d'en proposer une interprétation. Les modèles de type B permettent de segmenter les signes en utilisant l'ensemble de leur support et de corriger éventuellement les estimations de trajectoires issues du traitement d'image. Il est ensuite nécessaire d'utiliser des modèles de type A, afin de supprimer le contexte et de ne décrire que le noyau du signe. Ce sont la structure et les paramètres de ce noyau qui permettront de reconnaître le signe.

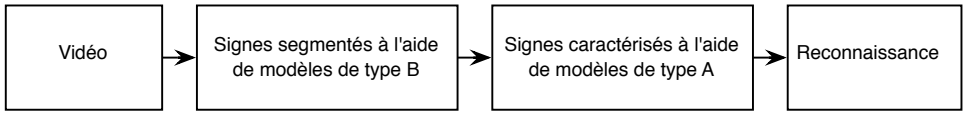


FIGURE 9.2 – Couplage entre plusieurs types de modèles pour la reconnaissance.

9.2.2 Génération d'énoncés signés en synthèse pure

Dans le cas de génération d'énoncés en synthèse pure, aucun exemple de signe issu de capture de mouvements n'est utilisé. Dans un premier temps, l'énoncé est décrit comme une suite de signes décrits à l'aide de modèles de types A et paramétrés à partir de valeurs P_i^A dépendant de règles sémantiques ou syntaxiques. A ce stade, tous les signes sont décrits hors contexte. Dans un second temps, les signes sont décrits cette fois, à l'aide de modèles de type B. On adjoint aux paramètres P_i^A, d'autres paramètres P_i^B dépendants de règles phonologiques et liés aux coarticulations, au style du signeur ... La description géométrique des signes est ensuite transformée en série d'instructions qui permettent d'animer le signeur virtuel.

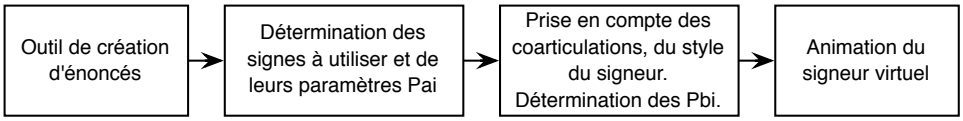


FIGURE 9.3 – Couplage de plusieurs modèles de signes pour effectuer de la synthèse pure.

9.2.3 Génération d'énoncés signés en utilisant des données de capture de mouvement

Contrairement au cas de figure que nous décrivions précédemment, nous disposons déjà ici, de réalisations de signes S_j acquises par capture de mouvement et déjà coarticulées, qu'il s'agit de réutiliser dans un autre contexte. Nous pouvons utiliser ces enregistrements pour déterminer les paramètres $P_i'^B$ correspondant aux signes S_j. La génération peut, comme précédemment, s'effectuer en paramétrant successivement les modèles de signes A et B, à la différence près que les descriptions des signes ne sont pas transformées directement en une série d'instructions pour le signeur virtuel. Au lieu de cette étape, on déforme spatialement et temporellement les réalisations de signes paramétrées par $P_i'^B$ de manière à ce que les nouveaux paramètres correspondent aux P_i^B que nous souhaitons utiliser pour paramétrer le signe dans le cadre de la synthèse de l'énoncé par le signeur virtuel.

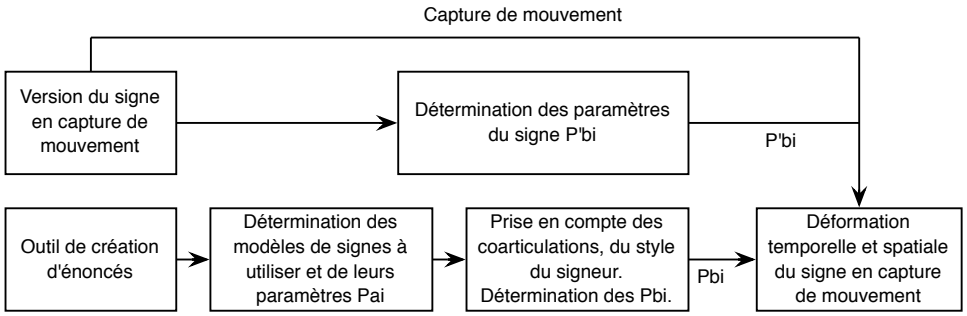


FIGURE 9.4 – Couplage de plusieurs modèles de signes pour la synthèse à partir de données de capture de mouvement.

9.2.4 Paramétrage de modèles à partir de traitement vidéo

Il est souvent délicat de déterminer les valeurs P_i^A et P_i^B nécessaires pour paramétrer les signes. Cette difficulté peut être liée à des lacunes dans les modèles sémantiques, syntaxiques, phonologiques ou prosodiques servant à la génération d'énoncés en LS ou d'un manque de temps pour effectuer l'initialisation des paramètres à la main. Dans ce cas, il peut être avantageux de réaliser l'énoncé en face d'une caméra et de déterminer les paramètres par traitement d'image. La capture vidéo de l'énoncé à synthétiser présente l'intérêt d'être à la fois peu coûteuse et de nécessiter un temps minime. Les paramètres déterminés à partir de la vidéo peuvent ensuite être réutilisés pour effectuer de la synthèse pure ou basée sur la capture de mouvement (cf. figures 9.3 et 9.4). Soulignons toutefois qu'une telle approche ne permettrait pas forcément d'accéder précisément à tous les paramètres P_i^B et $P^{B\prime}{}_i$ faisant intervenir la profondeur.

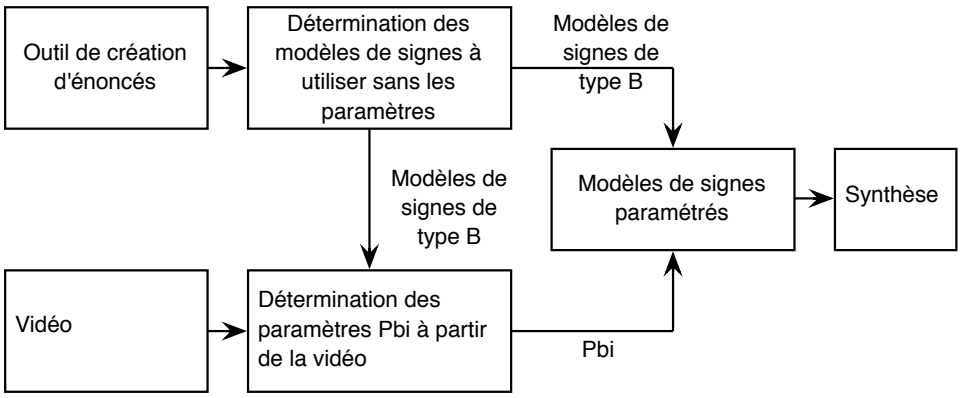


FIGURE 9.5 – Analyse de vidéo pour la synthèse.

9.3 Bilan

On l'aura compris au travers de ces perspectives, l'avenir du traitement automatique des LS est dans le croisement des savoirs de différentes disciplines. Le traitement bas niveau de vidéos en LS est si complexe qu'il nécessitera certainement de prendre en compte un certain nombre de considérations linguistiques permettant d'apporter des hypothèses simplificatrices, ou d'effectuer un traitement descendant de vidéos en LSF. La coopération entre linguistes et informaticiens permettra certainement d'aboutir à terme, à une interprétation de plus haut niveau des énoncés. Le couplage de plusieurs modèles permettra de prendre davantage en compte les différents niveaux de variabilité du signe pour une meilleure analyse et une meilleure synthèse d'énoncés en LS.

CHAPITRE 10

CONCLUSION

10.1 Nos principales contributions

Cette thèse multidisciplinaire nous a permis de proposer des modèles et des méthodes de traitement automatique de vidéos en LS. Nos contributions sont à la fois des apports théoriques et des applications dédiées au traitement automatique de la vidéo. Nous les énumérons rapidement dans les parties qui suivent, en les classant par domaines de recherche.

10.1.1 Suivi et reconstruction de la posture

Nous avons proposé une nouvelle méthode de suivi des différentes parties du corps d'un signeur combinant les avantages des filtres particulaires partitionnés et des images intégrales. Cette méthode prend en compte les problèmes d'occultations et de sortie de cadre des mains du signeur.
Une autre solution originale a été apportée au problème de désambigüisation des mains droite et gauche, nécessaire à une reconstruction cohérente des postures du signeur.
La corrélation entre les profondeurs des mains dominante et dominée a été mise à profit pour diminuer l'erreur d'estimation sur les profondeurs estimées par cinématique inverse.
Pour finir, l'utilisation de modèles de signes a permis d'aboutir à une amélioration de l'estimation de la position des mains, basée sur l'utilisation de modèles de signes.

10.1.2 Modèle de mouvement

Les différentes expérimentations que nous avons exposées démontrent empiriquement la pertinence de notion de *signe*, même si sa définition linguistique est difficile à formuler. Nous montrons qu'il est possible d'aboutir à une segmentation cohérente d'un énoncé en signes sans faire intervenir la compréhension de l'énoncé. Les résultats encourageants des expériences de segmentations basées uniquement sur le mouvement prouvent que la seule utilisation du mouvement peut permettre de déterminer le début et la fin de certains signes.
Nous proposons un modèle de description paramétrique des signes que nous avons approfondi davantage pour les mouvements balistiques et les mouvements balistiques répétés, qui représentent à eux seuls plus de la moitié des signes du lexique standard.
Finalement, notre modèle paramétrique du mouvement pour l'analyse est utilisé pour déterminer une mesure de similarité entre deux signes d'une même catégorie.

10.1.3 Applications

Les algorithmes de suivi et les modèles de signes sont enfin utilisés dans le cadre d'applications dédiées au traitement automatique de vidéos en LS.
La première permet d'effectuer une recherche de signe dans un énoncé en LS en utilisant uniquement l'information du mouvement. La deuxième a pour but d'aider un utilisateur dans la création d'une représentation graphique du signe en automatisant la fusion d'images d'une vidéo par transparence et le tracé des flèches figurant le mouvement impliqué dans le signe. La troisième permet d'assister un opérateur humain dans la tâche de segmentation d'un énoncé en signes. Cette application est également basée sur les modèles de mouvements.

10.2 Réponse à la problématique

Le but de ce travail était de déterminer *comment il était possible d'intégrer les règles phonologiques régissant le mouvement des signes standards de la LSF, dans un système de traitement automatique de vidéos d'énoncés en LSF*. Les différentes contributions présentées dans le paragraphe précédent montrent que ces règles peuvent être introduites à différentes étapes du traitement de vidéo :

- Dans le traitement bas niveau de la vidéo, ces règles permettent de désambigüiser les mains du signeur et de donner une meilleure estimation de la profondeur des mains.
- Les règles phonologiques peuvent naturellement être utilisées pour créer des modèles de mouvement destinés à l'analyse de vidéos.
- Les modèles de signes peuvent ensuite être utilisés pour corriger des défauts de suivi des mains.
- Ces modèles permettent également de déterminer des similarités entre signes, qui pourront être utilisées dans le cadre de recherche par le contenu dans une vidéo.
- Chaque patron de mouvement peut être associé à une représentation graphique normalisée. L'utilisation de règles sur la structure des mouvements permet de représenter la majorité des mouvements avec un petit nombre de primitives graphiques.
- Enfin, partant de l'observation qu'il n'est pas forcément nécessaire de comprendre un énoncé composé de signes standards pour le segmenter, nous proposons une méthode de segmentation utilisant la notion de patrons de mouvements. Segmenter un énoncé revient alors à en proposer un découpage phonologiquement cohérent.

Il reste maintenant à pousser plus loin l'analyse des différents patrons de mouvement pour en dresser une liste plus exhaustive et identifier tous les paramètres pour chacun des patrons. D'autres questions qui se posent sont relatives au statut des différents paramètres. A quels niveaux (sémantique,

syntaxique, prosodique ...) interviennent ils ? Quels sont leurs importances respectives dans le processus de reconnaissance des signes ? Certains sont ils de nature discrète ? Le style du signeur se trouve-t-il inscrit dans ces paramètres ?

10.3 Apport dans les autres domaines

Nos contributions concernent les langues des signes, mais seraient peut-être transposables à d'autres domaines. Nous évoquons dans les lignes qui suivent, plusieurs champs de recherche connexes au nôtre où notre démarche pourrait être mise à profit.

La synthèse d'énoncés en LS est naturellement l'un des domaines où il serait possible d'appliquer les modèles que nous proposons. Ce point est développé dans nos perspectives. Par extension, il serait certainement possible d'utiliser une démarche similaire pour effectuer une synthèse d'autres catégories de gestes car nous savons d'après les travaux de [GVKVDH98] que certaines la symétrie des mouvements est aussi fréquemment visible dans la gestualité coverbale.

Le fait d'utiliser des modèles du système à observer n'est pas nouveau dans le domaine du traitement automatique de la vidéo. Par contre, la manière dont nous utilisons les modèles de mouvement est peu répandue. L'approche serait transposable à différents types de mouvements. En plus des gestes coverbaux que nous évoquions précédemment, elle pourrait certainement être appliquée aux mouvements sportifs, aux mouvements musicaux ou à la danse. Nous avons d'ailleurs appliqué notre algorithme de suivi au suivi de mains du pianiste en marge de notre travail sur la LS [LAD09b].

10.4 Bilan

En trois ans, nous avons été les témoins de grand progrès dans le domaine du traitement automatique des Langues des Signes. Les traitements effectués sur les vidéos sont de plus en plus efficaces et permettent de traiter des données moins contraintes, tant au niveau du contenu des énoncés qu'au niveau des conditions d'acquisition. Cette dynamique dans laquelle nous nous inscrivons permet la mise au point de fonctionnalités de traitement de plus haut niveau pour le traitement d'énoncés réels, comme l'automatisation de la création d'image de signe, la reconnaissance, la segmentation ou l'analyse pour la génération.
Certains de ces outils comme Photosigne ou VAES sont déjà fonctionnels et peuvent être utilisés dans le cadre de l'enseignement des LS. D'autres comme les traducteurs automatique entre deux LS ou les bornes interactives entre un signeur réel et un signeur virtuel sont encore à l'état de recherche.

Espérons que tous ces outils pourront être utilisés par la communauté des signeurs pour enseigner, diffuser et faire vivre cette Langue qui nous lance sans cesse de nouveaux défis.

BIBLIOGRAPHIE

[AA01] S. Akyol and P. Alvarado. Finding relevant image content for mobile sign language recognition. In *Proceedings of the International Conference on Signal Processing, Pattern Recognition and Application*, pages 48–52, Rhodes, Greece, 2001. IEEE.

[AG98] M. Assan and K. Grobel. Video-based sign language recognition using hidden markov models. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 97–109, London, UK, 1998. Springer-Verlag.

[Alo06] J. Alon. *Spatiotemporal gesture segmentation*. PhD thesis, University of Boston, USA, 2006.

[AT04] A. Agarwal and B. Triggs. Learning to track 3d human motion from silhouettes. In *Proceedings of the twenty-first International Conference on Machine Learning*, pages 9–16, New York, USA, 2004. ACM.

[BA25] A. Bebian and R. Ambroise. *Mimographie, ou Essai d'écriture mimique, propre à régulariser le langage des sourds-muets*. Paris : L. Colas, 1825.

[Bac76] C. Backer. What's not on the other hand in american sign language. *Papers from the Twelfth Regional Meeting of the Chicago Linguistic Society*, 12 :24–32, 1976.

[Bat74] R. Battison. Phonological deletion in american sign language. *Sign Language Studies*, 5 :1–19, 1974.

[BBFV07] A. Braffort, L. Bolot, M. Filhol, and C. Verrecchia. Démonstrations d'elsi, la signeuse virtuelle du limsi. In *colloque Traitement Automatique des Langues des Signes, Atelier Traitement Automatique des Langues des Signes*, Toulouse, France, 2007.

[BE95] M. Brand and I. Essa. Causal analysis for visual gesture understanding. In *In Proceedings of AAAI Fall Symposium on Computational Models for Integrating Language and Vision*, Cambridge, Royaume-Uni, 1995.

[BEJ07] A. Bernard, F. Encrevé, and F. Jeggli. *L'interprétation en langue des signes*. Presses Universitaires de France, 2007.

[Bel03] R.E. Bellman. *Dynamic Programming*. Dover paperback edition, 2003.

[BEZre] P. Buehler, M. Everingham, and A. Zisserman. Employing signed tv broadcasts for automated learning of british sign language. In *International Conference on Language Resources and Evaluation, 4th Workshop on the Representation and Processing of Sign Languages : Corpora and Sign Language Technologies*, Malte, 2010 (à paraître).

[BH94] A.M. Baumberg and D.C. Hogg. Learning flexible models from image sequences. In *Proceedings of the third European Conference on Computer Vision*, volume 1, pages 299–308, Stocklholm, Suède, 1994. Springer-Verlag.

[Bin09] T. Binet, A. Simon. Peut-on enseigner la parole aux sourds-muets ? *L'Année psychologique*, 15 :373–396, 1909.

[BM98] C. Bregler and J. Malik. Video motion capture. In *Proceedings of SIGGRAPH (Special Interest Group in GRAPHics)*, Orlando, USA, 1998.

[BM00a] S. Belongie and J. Malik. Matching with shape contexts. In *Workshop on Content-Based Access of Image and Video Libraries*, Hilton Head, USA, 2000.

[BM00b] J. Brand and J.S. Mason. A comparative assessment of three approaches to pixel-level human skin-detection. In *15th International Conference on Pattern Recognition*, volume 1, pages 1056–1059, Barcelone, Espagne, 2000.

[BMGB09] M. Barnachon, B. Michoud, E. Guillou, and S. Bouakaz. Reconstruction géométrique par estimation de posture. In *ORASIS*, Trégastel, France, 2009.

[BOP97] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 994, 1997.

[Bos02] B. Bossard. Problèmes posés par la reconnaissance de gestes en langue des signes. In *Conférence annuelle sur le traitement automatique des langues naturelles, Recital*, volume 1, page 445, Nancy, France, 2002.

[Bou07] B. Boutet, D. Garcia. Compositionnalité morphophonétique de la langue des signes française (lsf) et exploration des relations structurales entre paramètres. *revue Traitement Automatique des Langues : Modélisation et traitement des Langues des Signes*, 48 :89–110, 2007.

[Bou09] L. Boutora. *Fondements historiques et implications théoriques d'une phonologie des langues des signes*. PhD thesis, Université Paris VIII, Vincennes - Saint Denis, 2009.

[BPSW70] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41 :164–171, 1970.

[Bra96] A. Braffort. *Reconnaissance et compréhension de geste, application à la langues des signes*. PhD thesis, UFR Sciences, LIMSI, Université Paris XI, 1996.

[Bra00] J. Bradbury. Linear predictive coding. http ://www.cs.queensu.ca/home/bradbury/pdf/lpc paper.pdf, 2000.

[Bre98] D. A Brentari. *Prosodic Model of Sign Language Phonology*. MIT Press, 1998.

[BSB05] A. O. Balan, L. Sigal, and M. J. Black. A quantitative evaluation of video-based 3d person tracking. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 349–356, Washington, USA, 2005. IEEE Computer Society.

[BW97] Aaron F. Bobick and Andrew D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(12) :1325–1337, 1997.

[BWK+04] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In Tom·s Pajdla and Jiri Matas, editors, *European Conference on Computer Vision*, volume 3021 of *Lecture Notes in Computer Science*, pages 390–401, Prague, République Tchèque, 2004. Springer.

[CADlT04] J. Cortadellas, J. Amat, and F. De la Torre. Robust normalization of silhouettes for recognition applications. *Pattern Recognition Letters*, 25 :591 – 601, 2004.

[Can86] F.J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6) :679–698, 1986.

[CBA+96] L.W. Campbell, D.A. Becker, A. Azarbayejani, A.F. Bobick, and A. Pentland. Invariant features for 3-d gesture recognition. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pages 157–162, Killington, USA, 1996. IEEE Computer Society.

[Cha02] R.E. Channon. *Signs are single segments : phonological representations and temporal sequencing in ASL and other sign languages*. PhD thesis, University of Maryland, Washington, USA, 2002.

[Cho57] N. Chomsky. *Structures syntaxiques*. Le seuil, 1957.

[CJM04] T. Chateau, F. Jurie, and R. Marc. Reconnaissance de gestes par vision monoculaire temps réel : application à la formation des chargés de manœuvres pour la conduite des ponts polaires. In *congrès francophone Reconnaissance des Formes et Intelligence Artificielle*, Toulouse, France, 2004.

[Com06] M. Companys. *Dictionnaire 1200 signes, la Langue des Signes Française*. Editions Monica Companys, 2006.

[CR99] T.J. Cham and J.M. Rehg. A multiple hypothesis approach to figure tracking. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 239–245, Fort Collins, USA, 1999. IEEE Computer Society.

[CSW95] Y. Cui, D.L. Swets, and J.J. Weng. Learning-based hand sign recognition using shoslif-m. In *International Conference on Computer Vision*, pages 631–636, Los Alamitos, USA, 1995. IEEE Computer Society.

[Cux00] C. Cuxac. *Les voies de l'iconicité*. Ophrys, 2000.

[Cux02] C. Cuxac. "langage et cognition" , rapport de fin de recherche. http ://www.irit.fr/LS-COLIN, 2002.

[DBGP96] E. Di Bernardo, L. Goncalves, and P. Perona. Monocular tracking of the human arm in 3d : real-time implementation and experiments. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 3, pages 622–626, Vienne, Autriche, 1996.

[DBR00] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition*, volume 2, pages 126–133, Hilton Head, USA, 2000. IEEE Computer Society.

[DC01] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *International Conference on Computer Vision*, pages 315–320, Vancouver, Canada, 2001. IEEE Computer Society.

[DD92] A.C. Downton and H. Drouet. Model-based image analysis for unconstrained human upper-body motion. In *International Conference on Image Processing and its Applications*, pages 274–277, Maastricht, Pays-Bas, 1992. IEEE Computer Society.

[DDR+05] M. Debar, M. Desvignes, B. Romaniuk, G. Bailly, and Y. Payan. Reconstruction par noyaux polynomiaux. In *Groupe de Recherche et d'Etudes du Traitement du Signal et des Images - Actes de Colloques*, volume 1, pages 1137–1140, Louvain-la-Neuve, Belgique, 2005.

[Del07] Y. Delaporte. *Dictionnaire étymologique et historique de la Langue des Signes Française - Origine et évolution de 1200 signes*. Editions du Fox, 2007.

[Der07] K.G. Derpanis. Integral image-based representations. http ://www.cse.yorku.ca/ kosta/CompVisNotes/integralrepresentations.pdf, 2007.

[DF99] Q. Delamarre and O. Faugeras. 3d articulated models and multi-view tracking with silhouettes. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 716–721, Washington, USA, 1999. IEEE Computer Society.

[DFS08] J.D. Durou, M. Falcone, and M. Sagona. Numerical Methods for Shape-from-shading : A New Survey with Benchmarks. *Computer Vision and Image Understanding*, 109(1) :22–43, 2008.

[DGTC08] J. Dalle, S. Gonzalez, and A. Thaumoux-Crozat. *Nouvelles formes lexicales et enseignement de la LSF, mémoire de projet tutoré*. PhD thesis, Université Paris VIII, Vincennes - Saint Denis, France, 2008.

[DL05] Patrice Dalle and Boris Lenseigne. Vision-based sign language processing using a predictive approach and linguistic knowledge . In *IAPR conference on Machine Vision Applications*, pages 510–513, Tsukuba Science City, Japon, 2005. The International Association for Pattern Recognition (IAPR).

[DM06] L. Ding and A.M. Martinez. Three-dimensional shape and motion reconstruction for the analysis of american sign language. In *Conference on Computer Vision and Pattern Recognition Workshop*, page 146, New York, USA, 2006.

[DP93] T. Darrell and A. Pentland. Space-time gesture. In *Conference on Computer Vision and Pattern Recognition*, pages 335–340, New York, USA, 1993. IEEE Computer Society.

[DSD+08] P. Dreuw, D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth, and H. Ney. Spoken language processing techniques for sign language recognition and translation. *Technology and Dissability*, 20(2) :121–133, 2008.

[DWT04] K.G. Derpanis, R.P. Wildes, and J.K. Tsotsos. Hand gesture recognition within a linguistics-based framework. In *Proceedings of the European Conference on Computer Vision*, pages 282–296, Prague, République Tchèque, 2004. Springer.

[Eco88] U. Eco. *Le Signe, histoire et analyse d'un concept, adapté de l'italien par Jean-Marie Klinkenberg*. Editions Labor, 1988.

[FB07] M. Filhol and A. Braffort. Description lexicale des signes, intérêts linguistiques d'un modèle géométrique à dépendances. *revue TAL : "Modélisation et traitement des langues des signes", direction P. Dalle et C. Cuxac*, 48,2, 2007.

[FF05] K. Forbes and E. Fiume. An efficient search algorithm for motion data using weighted pca. In *Proceedings of the Eurographics symposium on Computer animation*, pages 67–76, Los Angeles, California, 2005. Association for Computing Machinery.

[FGZ04] G. Fang, W. Gao, and D. Zhao. Large vocabulary sign language recognition based on fuzzy decision trees. *Transaction Systems, Man and Cybernetics*, 34(3) :305–314, 2004.

[Fil08] M. Filhol. *Modèle descriptif des signes pour un traitement automatique des langues des signes*. PhD thesis, Université Paris-11, Orsay, France, 2008.

[FLD08] M. Fontmarty, F. Lerasle, and P. Danès. Une stratégie hybride de filtrage particulaire pour le suivi de mouvement humain depuis un robot mobile. In *Reconnaissance de Formes et Intelligence Artificielle*, page 10p, Amiens, France, 2008.

[Fon08] M. Fontmarty. *Vision et filtrage particulaire pour le suivi tridimensionnel de mouvement humain. Applications à la Robotique*. PhD thesis, Université Paul Sabatier, Toulouse, France, 2008.

[For01] D.A. Forsyth. Shape from texture and integrability. In *International Conference on Computer Vision*, pages 447– 453, Vancouver, Canada, 2001. IEEE Computer Society.

[Fou07] C. Fournier. Le fournier signé, dictionnaire bilingue et informatisé lsf/français. CD édité par le CRDP, 2007.

[Fre14] F.N. Freeman. Experimental analysis of the writing movement. *Psychological Review Monographs Supplement 17*, pages 1–46, 1914.

[FSC93] M. Fanty, P. Schmid, and R. Cole. City name recognition over the telephone. In *International Conference of Acoustics, Speech, and Signal Processing*, pages 549–552, Minneapolis, Minnesota , USA, 1993. IEEE Computer Society.

[Gav99] D.M. Gavrila. The visual analysis of human movement : A survey. *Computer Vision and Image Understanding : CVIU*, 73(1) :82–98, 1999.

[GBM+07] B. Garcia, J.L. Brugeille, H. Mercier, P. Dalle, and A. Braffort. Projet ls-script : acte de la journée dans le cd-rom accompagnant le rapport final du projet, 2007.

[GCG+96] P.H. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan. Multi-modal system for locating heads and faces. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pages 88–93, Killington, Vermont, USA, 1996. IEEE Computer Society.

[GCH99] A. Garg, I. Cohen, and T.S. Huang. Adaptive learning algorithm for svm applied to feature tracking. In *International Conference on Information Intelligence and Systems*, Washington, USA, 1999. IEEE Computer Society.

[Gia08] F. Gianni. *Suivi de parties du corps pour l'interprétation de gestes de communication à partir de séquences monoculaires*. PhD thesis, Université Paul Sabatier, Toulouse, France, 2008.

[Gib02] S. Gibet. Modèles d'analyse-synthèse du mouvement, habilitation à diriger des recherches, 2002.

[Gor93] D.J. Gordon, N.J. Salmond. Novel approach to nonlinear / non-gaussian bayesian state estimation. *Processings for Radar and Signal Processing*, 1400 :107–113, 1993.

[GPS+06] J. Gall, J. Potthoff, C. Schnoerr, B. Rosenhahn, and H.P. Seidel. Interacting and annealing particle filters : Mathematics and a recipe for applications. Technical report, institut Max-Planck, Mayence, Allemagne, 2006.

[GVKVDH98] I. Gijn (Van), S. Kita, and H. Van Der Hulst. The non-linguistic status of the symmetry condition in signed languages : Evidence from a comparison of signs and spontaneous co-speech gestures, 1998.

[GW06] D. Guo and X. Wang. Quasi-monte carlo filtering in nonlinear dynamic systems. *Transactions on Signal Processing*, 54(6) :2087–2098, 2006.

[HCGM06] A. Heloir, N. Courty, S. Gibet, and F. Multon. Alignement temporel de séquences gestueles communicatives, June 2006.

[HCZ08] M. Hruz, P. Campr, and M. Zelezny. Semi-automatic Annotation of Sign Language Corpora. In *international conference on Language Resources and Evaluation*, pages 78–81, Marrakech, Maroc, 2008.

[HGO96] H. Hienz, K. Grobel, and G. Offner. Real-time hand-arm motion analysis using a single video camera. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 323–327, 1996.

[HH98] C.L. Huang and W.Y. Huang. Sign language recognition using model-based tracking and a 3d hopfield neural-network. *Machine Vision and Applications*, 10(5-6) :292–307, 1998.

[HHD98] I. Haritaoglu, D. Harwood, and L.S. Davis. Ghost : A human body part labeling system using silhouettes. In *International Conference on Pattern Recognition*, volume 1, pages 77–82, Brisbane, Australie, 1998.

[Hje43] L. Hjelmslev. *Prolégomènes à une théorie du langage*. Editions de minuit, 1943.

[HLA09] A. Hadjakos and F. Lefebvre-Albaret. Three methods for pianist hand assignment. In *6th Sound and Music Computing Conference*, pages 321–326, Porto, Portugal, 2009.

[HLM04] N. Habili, C.C. Lim, and A. Moini. Segmentation of the face and hands in sign language video sequences using color and motion cues. *Transactions on circuits and systems for video technology*, 14(8) :1086–1097, 2004.

[HP91] B. Horowitz and A. Pentland. Recovery of non-rigid motion and structure. In *Computer Vision and Pattern Recognition*, pages 325–330, Hawaii, 1991. IEEE Computer Society Conference.

[HREAU08] J.L. Hernandez-Rebollar, E.I. Elsakay, and J.D. Alanis-Urquieta. Accelespell, a gestural interactive game to learn and practice finger spelling. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, pages 189–190, Chania, Crete, Grèce, 2008. Association for Computing Machinery.

[HSA95] T. Heap, O.A. Site, and Trumpington A.S. Real-time hand tracking and gesture recognition using smart snakes, 1995.

[Hu62] M.K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8 :179–187, 1962.

[HVd93] H. Hulst (Van der). Units in the analysis of signs. *Phonology*, 10 :209–241, 1993.

[HVM96] N. Harte, S.V. Vaseghi, and B. Milner. Dynamic features for segmental speech recognition. In *International Conference on Spoken Language Processing*, pages 933–936, Philadelphie, USA, 1996. IEEE Computer Society.

[HYDD05] B. Han, C. Yang, R. Duraiswami, and L. Davis. Bayesian filtering and integral image for visual tracking. In *Workshop on Image Analysis for Multimedia Interactive Services*, Montreux, Suisse, 2005.

[IB98a] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29 :5–28, 1998.

[IB98b] M. Isard and A. Blake. Icondensation : Unifying low-level and high-level tracking in a stochastic framework. In *European Conference on Computer Vision*, pages 893–908, Fribourg, Allemagne, 1998. Springer.

[ILI98] K. Imagawa, S. Lu, and S. Igi. Color-based hands tracking system for sign language recognition. In *Proceedings of the 3rd. International Conference on Face Gesture Recognition*, pages 462–467, Nara, Japon, 1998. IEEE Computer Society.

[JB03] S. Jehan-Besson. *Suivi de parties du corps pour l'interprétation de gestes de communication à partir de séquence monoculaire*. PhD thesis, Université Paul Sabatier, Toulouse, France, 2003.

[JBY96] S.X. Ju, M.J. Black, and Y. Yacoob. Cardboard people : a parameterized model of articulated image motion. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 38–44, Killington, Vermont, USA, 1996. IEEE Computer Society.

[Jes22] O. Jespersen. *Language. Its Nature, Development and Origin*. Allen Unwin, 1922.

[JJC02] D.S. Jang, S.W. Jang, and H.I. Choi. 2d human body tracking with structural kalman filter. *Pattern Recognition*, 35(10) :2041–2049, 2002.

[JM09] A. Just and S. Marcel. A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition. *Computer Vision and Image Understanding*, 113(4) :532–543, 2009.

[Joh08] R. Johnson. Atelier sur la notation phonétique des langues des signes. In *Sans actes*, Orsay, France, 2008.

[Jor90] M.I. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. *Artificial neural networks : concept learning*, pages 112–127, 1990.

[Jou95] B. Jouison, P Garcia. *Ecrits sur la langue des signes française*. L'Harmattan, 1995.

[KC01] I.C. Kim and S.I. Chien. Analysis of 3d hand trajectory gestures using stroke-based composite hidden markov models. *Applied Intelligence*, 15(2) :131–143, 2001.

[Ken88] A. Kendon. How gesture can become like words. *crosscultural perspectives in nonverbal communication*, pages 131–141, 1988.

[KH97] T. Kobayashi and S. Haruyama. Partly-hidden markov model and its application to gesture recognition. *International Conference on Acoustics, Speech, and Signal Processing*, 4 :3081, 1997.

[Kit96] G. Kitagawa. Monte-carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1) :1–25, 1996.

[KJB96] J.S. Kim, W. Jang, and Z.N. Bien. A dynamic gesture recognition system for the korean sign language (ksl). *Systems, Man, and Cybernetics (B)*, 26(2) :354–359, 1996.

[Kub07] D. Kubrak. *Etude de l'hybridation d'un récepteur GPS avec des capteurs bas-coûts pour la navigation personnelle en milieu urbain*. PhD thesis, ENST - COMELEC Communication et Electronique, Paris, France, 2007.

[KVD06] S. Knoop, S. Vacek, and R. Dillmann. Sensor fusion for 3d human body tracking with an articulated 3d body model. In *International Conference on Robotics and Automation*, Orlando, Floride, USA, 2006. IEEE Computer Society.

[KvGvdH98] S. Kita, I. van Gijn, and H. van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. *Lecture notes in computer science*, 1371 :23–35, 1998.

[L.00] Ann J . Peng L. Optimality theory and opposed handshapes in taiwan sign language. *University of Rochester Working Papers in the Language Sciences*, 1 :173–194, 2000.

[Lab74] F.C. Laban, R. Lawrence. *Effort : Economy in body movement*. Plays, Inc, Boston, 1974.

[LAD09a] F. Lefebvre-Albaret and P. Dalle. Analyse de vidéos en langue des signes : méthodes et stratégies. In *ORASIS, Congrès des jeunes chercheurs en vision par ordinateur*, Trégastel, France, 2009.

[LAD09b] F. Lefebvre-Albaret and P. Dalle. Article sur le suivi des mains du pianiste. In *ORASIS09*, 2009.

[LAD09c] F. Lefebvre-Albaret and P. Dalle. Body posture estimation in a sign language video. In *International Gesture Workshop*, Bielefeld, Allemagne, 2009. Kirsten Bergmann, Hendrik Buschmeier (Eds.).

[LADD$^+$re] F. Lefebvre-Albaret, P. Dalle, J. Dalle, J.F. Piquet, S. Dalle-Nazébi, P. Gache, and A. Bacci. Analyse des langues des signes, démarche de conception pluridisciplinaire d'outils d'analyse de discours en langues des signes. *revue Technique et Sciences Informatiques, numéro spécial : l'informatique à l'interface de l'activité humaine et sociale*, 2010 (à paraître).

[LAGD08] F. Lefebvre-Albaret, F. Gianni, and P. Dalle. Toward a computer-aided sign segmentation. In *Language Resources and Evaluation Conference (LREC)*, Marrakech, Maroc, 2008. European Language Resources Association.

[Leb98] T. Lebourque. *Spécification et génération de gestes naturels. Application à la Langue des Signes Française*. PhD thesis, Université Paris XI Sud (LIMSI-CNRS), Orsay, France, 1998.

[LJ86] S.K. Liddell and R.E. Johnson. American sign language compound formation processes, lexicalization, and phonological remnants. *Natural Language Linguistic Theory*, 4, 1986.

[LJ90] S.K. Liddell and R.E. Johnson. American sign language : the phonological base. *Sign Language Studies*, 64, 1990.

[LO98] R.H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. *Conference on Automatic Face and Gesture Recognition*, page 558, 1998.

[Los00] O. Losson. *Synthèse du geste communicatif. Application à la langue des signes française*. PhD thesis, université de Lille, France, 2000.

[LPK+97] C.S. Lee, G. Park, J.S. Kim, Z. Bien, W. Jang, and Kim S.K. Real-time recognition system of korean sign language based on elementary components. In *Proceedings of the 6th International Conference on Fuzzy Systems*, pages 1463–1468. IEEE Computer Society, 1997.

[LR03] J. Le Roux. Modèles de markov cachés, 2003.

[LW07] P.H. Li and H.J. Wang. Object tracking with particle filter using color information. In *MIRAGE07, Computer Vision / Computer Graphics Collaboration Techniques and Applications*, pages 534–541, INRIA Rocquencourt, France, 2007.

[Mar70] A. Martinet. *Eléments de linguistique générale*. A.Colin, 1970.

[Mau09] T.H.H. Maung. Real-time hand tracking and gesture recognition system using neural networks. *World Academy of Science, Engineering and Technology*, pages 466–470, 2009.

[May79] P.S. Maybeck. Stochastic models, estimation and control. *Mathematics in Science and Engineering*, 1, 1979.

[MB99] S. Marcel and O. Bernier. Hand posture recognition in a body-face centered space. In *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pages 97–100, London, UK, 1999. Springer-Verlag.

[MBVC00] S. Marcel, O. Bernier, J.E. Viallet, and D. Collobert. Hand gesture recognition using input-output hidden markov models. *International Conference on Automatic Face and Gesture Recognition*, page 456, 2000.

[McN05] D. McNeill. *Gesture and Thought*. University of Chicago Press, 2005.

[Mer07] H. Mercier. *Modélisation et suivi des déformations faciales*. PhD thesis, Université Paul Sabatier, Toulouse, France, 2007.

[Meu08] L. Meurant. *Anaphore en langue des signes française de Belgique (LSFB) : morphologie, syntaxe, énonciation*. Presses Universitaires de Rennes / Presses Universitaires de Namur, 2008.

[MG01] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3) :231–268, 2001.

[MGW00] J. Ma, W. Gao, and R. Wang. A parallel multistream model for integration of sign language recognition and lip motion. In *Proceedings of the Third International Conference on Advances in Multimodal Interfaces*, pages 582–589, London, UK, 2000. Springer-Verlag.

[MGWW00] J. Ma, W. Gao, J. Wu, and C. Wang. A continuous chinese sign language recognition system. In *Fourth International Conference on Automatic Face and Gesture Recognition*, page 428, Washington, USA, 2000. IEEE Computer Society.

[MHM10] T. Matsubara, S.H. Hyon, and J. Morimoto. Learning stylistic dynamic movement primitives from multiple demonstrations. In *International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, 2010. IEEE.

[MHTAM07] Q. Munib, M. Habeeb, B. Takruri, and H.A. Al-Malik. American sign language (asl) recognition based on hough transform and neural networks. *Expert Systems with Applications*, 32(1) :24–37, 2007.

[MI00] John Maccormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conference on Computer Vision*, Dublin, Irlande, 2000.

[MM02] G. Mori and J. Malik. Estimating Human Body Configurations Using Shape Context Matching. In *Proceedings of the 7th European Conference on Computer Vision*, volume 3, Copenhague, Danemark, 2002. Springer.

[MOB06] A.S. Micilotta, E.J. Ong, and R. Bowden. Real-time upper body detection and 3d pose estimation in monoscopic images. In *European Conference on Computer Vision*, volume 3, pages 139–150, Graz, Autriche, 2006. Springer.

[Moo86] B. Moody. *La langue des signes. Dictionnaire bilingue élémentaire, Tome 1, 2 et 3*. IVT, 1986.

[MP88] W.S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Neurocomputing : foundations of research*, pages 15–27, 1988.

[MSL01] J.B. Martinkauppi, M.N. Soriano, and M.H. Laaksonen. Behavior of skin color under varying illumination seen by different cameras at different color spaces. *SPIE proceedings series*, 4301 :102–112, 2001.

[MT91] K. Murakami and H. Taguchi. Gesture recognition using recurrent neural networks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 237–242, Nouvelle-Orleans, USA, 1991. Association for Computing Machinery.

[MT93] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *Transactions on Pattern Analysis and Machine Intelligence*, 15(6) :580–591, 1993.

[MVGD97a] B. Moody, A. Vourch, M. Girot, and A.C. Dufour. *la langue des signes, dictionnaire bilingue LSF / Français*. IVT, 1997.

[MVGD97b] B. Moody, A. Vourch, M. Girot, and A.C. Dufour. *la langue des signes, histoire et grammaire*. IVT, 1997.

[NB07] P. Noriega and O. Bernier. Multicues 3D Monocular Upper Body Tracking using Constrained Belief Propagation. In *British Machine Vision Conference*, Warwick, Royaume-Uni, 2007.

[Nor07] P. Noriega. *Modèle du corps humain pour le suivi de gestes en monoculaire*. PhD thesis, Université Pierre et Parie Curie, Paris, France, 2007.

[NW96] Y. Nam and K.Y. Wohn. Recognition of Space-Time Hand-Gesture using Hidden Markov Model. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pages 51–58, Hong Kong, 1996. Association for Computing Machinery.

[OR05] S.C.W. Ong and S. Ranganath. Automatic sign language analysis : A survey and the future beyond lexical meaning. *Transactions on Pattern Analysis and Machine Intelligence*, 27(6) :873–891, 2005.

[OR08] A.M. Othman and M.H. Riadh. Speech recognition using neural networks. *World Academy of Science, Engineering and Technology*, 38 :253–258, 2008.

[Par03] A.S. Parashar. *Representation and interpretation of manual and non-manual information for automated American Sign Language recognition*. PhD thesis, Université de Floride du sud, USA, 2003.

[Per92] D.M. Perlmutter. Sonority and syllable structure in american sign language. *Linguistic Inquiry*, 23, 3 :407–442, 1992.

[PF01] R. Plankers and P. Fua. Articulated soft objects for video-based body modeling. In *International Conference on Computer Vision*, pages 394–401, Vancouver, Canada, 2001. IEEE Computer Society.

[PF04] E. Prados and O. Faugeras. Unifying approaches and removing unrealistic assumptions in shape from shading : Mathematics can help. In Jiri Matas Tomas Pajdla, editor, *Proceedings of the 8th European Conference on Computer Vision*, volume 3024, pages 141–154, Prague, République Tchèque, 2004. Springer.

[PHD10] J. Piater, T. Hoyoux, and W. Du. Video analysis for continuous sign language recognition. In *4th Workshop on the Representation and Processing of Sign Languages : Corpora and Sign Language Technologies*, La Valetta, Malte, 2010.

[PNN94] R. Polana, R. Nelson, and A. Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, Austine, Texas, USA, 1994. IEEE Computer Society.

[PRC00] V. Pavlovic, J.M. Rehg, and T. Cham. A dynamic bayesian network approach to tracking using learned switching dynamic models. In *Proceedings of the International Workshop on Hybrid Systems*, Pittsburgh, Pennsylvanie, USA, 2000. Springer.

[Pri89] S. et Zienert H. Prillwitz. Hamburg notation system for sign language : Development of a sign writing with computer application. *Current trends in European Sign Language Research*, 1989.

[PSH97] V.I. Pavlovic, R. Sharma, and T.S. Huang. Visual interpretation of hand gestures for human-computer interaction : A review. *Transactions on Pattern Analysis and Machine Intelligence*, 19 :677–695, 1997.

[PVD03] G.S. Paul, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *International Conference on Computer Vision*, pages 750–757, Beijing, Chine, 2003. IEEE Computer Society.

[Rab89] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, 2, pages 257–286. IEEE Computer Society, 1989.

[RMR04] T.J. Roberts, S.J. McKenna, and I.W. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. *European Conference on Computer Vision*, 2004.

[RS00] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *Computer Vision and Pattern Recognition*, volume 2, pages 721–727, Hilton Head, Caroline du sud, USA, 2000.

[Sal03] M.A. Sallandre. *Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité*. PhD thesis, Université Paris VIII, Vincennes / Saint-Denis, France, 2003.

[San89] W. Sandler. *Phonological representation of the sign : linearity and nonlinearity in American Sign Language*. Dordrecht : Foris, 1989.

[Sau16] F. de Saussure. *Cours de linguistique générale*. Payot, 1916.

[SBS02] H. Sidenbladh, M.J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *European Conference on Computer Vision*, pages 784–800, Copenhague, Danemark, 2002.

[SC92] J. Shen and S. Castan. An optimal linear operator for step edge detection. In *Computer Vision, Graphical Models and Image Processing*, volume 54, 2, pages 112–133, Orlando, USA, 1992. Academic Press, Inc.

[SCC78] W.C. Stokoe, D. Casterline, and C. Croneberg. *A dictionary of American Sign Language on Linguistic principles*. Linstok Press, 1978.

[SFBL98] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin : a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26 :322–330, 1998.

[SG99] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252, Fort Collins, Colorado, USA, 1999. IEEE Computer Society Press.

[SG00] J. Sherrah and S. Gong. Resolving visual uncertainty and occlusion through probabilistic reasoning. In *Proceedings of the British Machine Vision Conference*, pages 252–261, Bristol, Royaume-Uni, 2000.

[SHJ94] J. Schlenzig, E. Hunter, and R. Jain. Vision based hand gesture interpretation using recursive estimation. In *28th Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1267 –1271. IEEE Computer Society Press, 1994.

[SK08] J.W. Sung and D.J. Kim. Pose-robust facial expression recognition using view-based 2d+3d aam. *Transactions on Systems, Man, and Cybernetics (A)*, 38(4) :852–866, 2008.

[SP95] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, Zurich, Suisse, 1995.

[SSA92] H. Sagawa, H. Sakou, and M. Abe. Sign language translation system using continuous dp matching. In *Proceedings of IAPR Workshop on Machine Vision Applications*, pages 339–342, Minato-ku, Tokyo, Japon, 1992.

[ST99] H. Sagawa and M. Takeuchi. A method for analyzing spatial relationships between words in sign language recognition. In *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pages 197–209, Gif-sur-Yvette, France, 1999. Springer-Verlag.

[ST00] H. Sagawa and M. Takeuchi. A method for recognizing a sequence of sign language words represented in a japanese sign language sentence. In *Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition*, page 434, Washington, USA, 2000. IEEE Computer Society.

[ST03] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *Computer Vision and Pattern Recognition*, pages 69–76, Madison, Wisconsin, USA, 2003. IEEE Computer Society Press.

[Sta76] N.L. Stauffer. Distance determining and automatic focusing apparatus. In *US Patent*, 1976.

[STO98] H. Sagawa, M. Takeuchi, and M. Ohki. Methods to describe and recognize sign language based on gesture components represented by symbols and numerical values. *Knowledge-Based Systems*, 10(5) :287 – 294, 1998.

[Su00] M.C. Su. A fuzzy rule-based approach to spatio-temporal hand gesture recognition. *Transactions on Systems, Man, and Cybernetics (C)*, 30(2) :276–281, 2000.

[Sut08] V. Sutton. Signwriting site. “ http ://www.signwriting.org/”, 2008.

[SVD03] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *International Conference on Computer Vision*, page 750, Beijing, Chine, 2003. IEEE Computer Society Press.

[SWP98] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *Transactions on Pattern Analysis and Machine Intelligence*, 20(12) :1371–1375, 1998.

[TA00] J.C. Terrillon and S. Akamatsu. Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images. In *12th Conference on Vision Interface*, pages 180–187, Montréal, Quebec, Canada, 2000.

[Tip00] M.E. Tipping. The relevance vector machine. In *Neural Information Processing Systems*, pages 652–658, 2000.

[TK92] C. Tomasi and T. Kanade. Shape and motion from image streams : a factorization method - full report on the orthographic case. *International Journal of Computer Vision*, 1992.

[TPNY02] J.C. Terrillon, A. Pilpre, Y. Niwa, and K. Yamamoto. Robust face detection and hand posture recognition in color images for human-machine interaction. *International Conference on Pattern Recognition*, 1 :10204, 2002.

[TvdM01] J. Triesch and C. von der Malsburg. A system for person-independent hand posture recognition against complex backgrounds. *Transactions on Pattern Analysis and Machine Intelligence*, 23 :1449–1453, 2001.

[TW88] S. Tamura and A. Waibel. Noise reduction using connectionist models. In *International Conference on Acoustics, Speech, and Signal Processing*, New York, USA, 1988.

[Uye97] L. Uyechi. *The Geometry of Visual Phonology*. CSLI Publications, Université de Stanford, Royaume-Uni, 1997.

[Vat08] R.D. Vatuvu. *Acquisition temps réel de la gestuelle humaine pour l’interaction en réalité virtuelle*. PhD thesis, Université de Lille, France, 2008.

[Vie00] E. Viel. *La marche humaine, la course et le saut : biomécanique, explorations, normes et dysfonctionnements*. Masson, Paris, France, 2000.

[VJ01] P.A. Viola and M.J. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, volume 1, pages 511–518, Hawaii, 2001.

[VM97] C. Vogler and D. Metaxas. Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. In *International Conference on Systems, Man and Cybernetics*, pages 156–161, Orlando, Floride, USA, 1997. IEEE Computer Society.

[VM98] C. Vogler and D. Metaxas. Asl recognition based on a coupling between hmms and 3d motion analysis. In *Sixth International Conference on Computer Vision*, page 363, Bombay, Inde, 1998. IEEE Computer Society.

[VM99a] C. Vogler and D. Metaxas. Parallel hidden markov models for american sign language recognition. In *International Conference on Computer Vision*, pages 116–122, Kerkyra, Corfu, Grèce, 1999. IEEE Computer Society.

[VM99b] C. Vogler and D. Metaxas. Toward scalability in asl recognition : Breaking down signs into phonemes. In *International Gesture Workshop*, page 211, Gif-sur-Yvette, France, 1999. Springer.

[Vog03] C.P. Vogler. *American sign language recognition : reducing the complexity of the task with phoneme-based modeling and parallel hidden markov models*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA, 2003. Supervisor-Dimitris N. Metaxas.

[VSM00] C. Vogler, H. Sun, and D. Metaxas. A framework for motion recognition with applications to american sign language and gait recognition. In *Proceedings of the Workshop on Human Motion*, page 33, Austin, Texas, USA, 2000. IEEE Computer Society.

[WADP96] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder : real-time tracking of the human body. In *Second International Conference on Automatic Face and Gesture Recognition*, pages 51–56, Killington, Vermont, USA, 1996. IEEE Computer Society.

[WB99] A.D. Wilson and A.F. Bobick. Parametric hidden markov models for gesture recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 21 :884–900, 1999.

[WHH+90] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. Phoneme recognition using time-delay neural networks. *Readings in speech recognition*, pages 393–404, 1990.

[WK95] M.B. Waldron and S. Kim. Isolated asl sign recognition system for deaf persons. *Transactions on Rehabilitation Engineering*, 1995.

[WN97] S. Wachter and H.-H. Nagel. Tracking of persons in monocular image sequences. In *Nonrigid and Articulated Motion Workshop*, pages 2–9, Porto Rico, 1997. IEEE Computer Society.

[WS03] J.J. Wang and S. Singh. Video analysis of human dynamics, a survey. *Real Time Imaging*, 9 :321–346, 2003.

[WS06] H. Wang and K. Shindler. Effective appearance model and similarity measure for particle filtering and visual tracking. *European Conference on Computer Vision*, 3953 :606–618, 2006.

[WSG02] C. Wang, S. Shan, and W. Gao. An approach based on phonemes to large vocabulary chinese sign language recognition. In *Proceedings of the Fifth International Conference on Automatic Face and Gesture Recognition*, page 411, Washington, USA, 2002. IEEE Computer Society.

[YA98] M.H. Yang and N. Ahuja. Extracting gestural motion trajectories. In *International Conference on Automatic Face and Gesture Recognition*, pages 10–15, Nara, Japon, 1998. IEEE Computer Society.

[YAT02] M.H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 24 :1061–1074, 2002.

[YC00] M. Yeasin and S. Chaudhuri. Development of an automated image processing system for kinematic analysis of human gait. *Real-Time Imaging*, 6(1) :55–67, 2000.

[YCW+08] Fang Y., J. Cheng, J. Wang, K. Wang, J. Liu, and H. Lu. Hand posture recognition with co-training. In *19th International Conference on Pattern Recognition*, pages 1–4, Tampa, Floride, USA, 2008. IEEE Computer Society.

[YTK+04] J. Yang, R.M. Timothy, H. Kim, J.S. Arora, and K. Abdel-Malek. Multi-objective Optimization for Upper Body Posture Prediction. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, New York, USA, 2004.

[Zha94] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2) :119–152, 1994.

[ZK05] J. Zieren and K.F. Kraiss. Robust person-independent visual sign language recognition. In *Iberian Conference on Pattern Recognition and Image Analysis*, volume 1, page 520, Estoril,Portugal, 2005.

ANNEXE I

LES FILTRES PARTICULAIRES

Nous souhaitons, dans cette annexe, présenter brièvement les différentes catégories de filtres particulaires qu'il est possible de mettre en œuvre dans le cadre de suivi de cibles mobiles. Le principe de tels filtres stochastiques est de maintenir à chaque instant plusieurs hypothèses sur l'état du système et de faire évoluer ces hypothèses en prenant en compte les connaissances *a priori* sur le système et les observations.

I.1 Principe général des filtres particulaires

L'état du système observé peut être décrit à chaque instant t par un vecteur d'état x_t appartenant à l'espace d'état E. On dispose par ailleurs à chaque instant d'un vecteur d'observations z_t appartenant à l'espace d'observation F. On suppose connue la distribution $p_0(x_0)$ du vecteur d'état à l'instant initial (Si cette distribution n'est pas connue, il est possible de choisir une distribution couvrant au mieux l'espace du vecteur d'état. On note en général que l'influence de la distribution initiale est de moins en moins importante lorsque t augmente). Une relation connue *a priori* $p(x_t|x_{t-1})$ permet de déterminer l'évolution du système. Une autre relation $p(z_t|x_t)$ indique la probabilité de l'observation z_t lorsqu'on est en présence de l'état x_t.

Le système est donc entièrement décrit par :

- La distribution $p_0(x_0)$ du vecteur d'état à l'instant initial,
- La dynamique $p(x_t|x_{t-1})$ du système,
- Le lien $p(z_t|x_t)$ entre l'observation et l'état du système.

L'objectif du filtrage particulaire est d'estimer à chaque instant t la distribution *a posteriori* d'état du système en utilisant l'ensemble des observations entre l'instant initial et l'instant t $(z_0, z_1 \dots z_{t-1}, z_t)$.

I.2 L'utilisation de particules

Dans certains cas, il est possible d'effectuer des hypothèses sur la forme des distributions utilisées dans le filtrage. Par exemple, dans le filtre de Kalman [May79], on suppose que la distribution initiale est gaussienne. On émet aussi l'hypothèse que la dynamique du système et le lien entre le vecteur d'état et le vecteur de mesure sont modélisés par des fonctions linéaires entachées d'un bruit

gaussien. Toutes les distributions utilisées dans de tels filtres sont donc gaussiennes et peuvent être caractérisées à l'aide de leur moyenne μ et de leur variance σ.

Dans de nombreux cas comme les gestes de la LSF, des hypothèses de distributions gaussiennes paraissent trop fortes et non réalistes. Si les distributions $p(x)$ sont quelconques, il est possible de les approximer par une distribution ponctuelle $\hat{p}(x)$ de N vecteurs d'états $x(i)$ pondérés par des poids $w(i)$. On distingue les différentes particules du nuage grâce à leur indice i. $\delta_{x(i)}$ représente une distribution de Dirac centrée en x(i).

$$p(x) \simeq \hat{p}(x) = \sum_{i=0}^{N} w(i).\delta_{x(i)}(x)$$

On peut alors simuler une variable aléatoire d'après la loi $p(x)$ donnant la probabilité de tirer une particule $x(i)$ avec un poids $w(i)$. Cet échantillonnage peut faire appel à une fonction d'importance $q(x)$ définie au moins sur l'espace E de manière à réaliser un *échantillonnage par importance*. Cet échantillonnage se déroule de la manière suivante[1] :

1. $x(i) \sim q(x)$ 2. $w'(i) \sim p(x(i))/q(x(i))$ 3. $w(i) = w'(i)/\sum_{i=0}^{N} w'(i)$

1. On effectue un certain nombre de tirages d'états d'après $q(x)$.
2. Les poids w'(i) sont calculés en tenant compte de la distribution de x et de la fonction d'importance.
3. Les poids w'(i) sont ensuite normalisés de manière à obtenir les pondérations des particules w(i).

Deux cas particuliers sont à distinguer :

- Si $q(x) = p(x)$, toutes les particules auront des poids identiques (fig. I.1 b).
- Si $q(x) = Cste$, les particules sont réparties aléatoirement dans l'espace E (fig. I.1 c).

Les particules permettent d'approximer des intégrales impliquant la distribiution $p(x)$ en utilisant l'approximation suivante :

$$\int_E f(x).p(x).dx \simeq \sum_{i=0}^{K} f(x(i)).w(i)$$

[1] La notation $v \sim D$ signifie que la valeur v a été générée d'après la distribution D

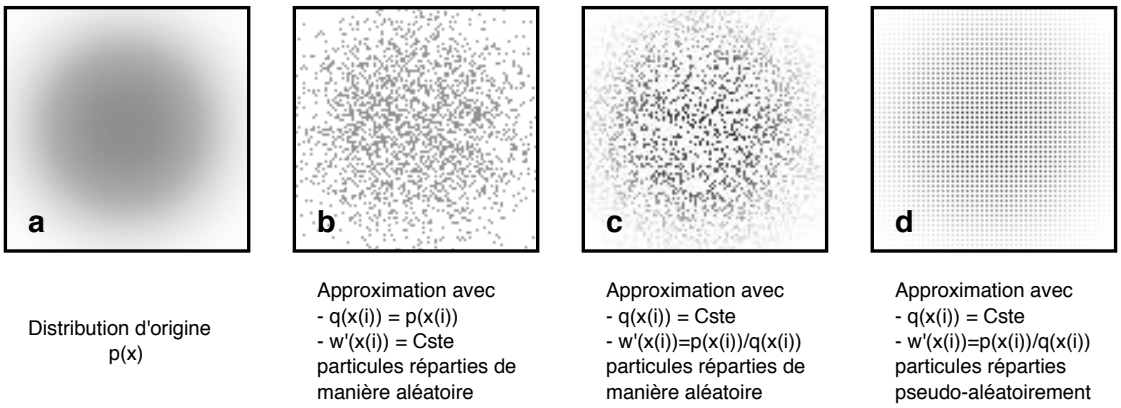


FIGURE I.1 – Approximation d'une distribution gaussienne à l'aide d'un nuage de particules

I.3 Filtrages particulaires simples

I.3.1 Propagation et mise à jour du poids des particules

Les premiers filtres particulaires sont mis au point indépendamment par [Gor93] et [IB98a]. Dans le cadre de vison par ordinateur, Isard et Blake nomment ces méthodes CONDENSATION (pour Conditional DENSity propagATION) tandis que Gordon et Salmond les nomment SIR (Sequential Importance Resampling).
Dans l'algorithme SIR, la distribution des vecteurs d'états $p(x)$ est approximée par un nuage de N vecteurs d'états $x(i)$ pondérés par des poids $w(i)$.

$$p(x) \simeq \sum_{i=0}^{N} w(i).\delta_{x(i)}(x)$$

A l'instant initial, l'état est approximé selon $p_0(x_0)$. Les auteurs de CONDENSATION utilisent la dynamique du système comme fonction d'importance. Cela revient à propager chaque particule en utilisant la dynamique du système $p(x_t|x_{t-1})$ et à mettre ensuite à jour leur poid à l'aide de la relation état-observation $p(z|x)$.
La mise à jour des poids est donc calculée grâcé à la relation suivante :

$$w'_t(i) = w_{t-1}(i).p(z|x_t(i))$$

Les poids des particules sont ensuite normalisés de manière que la masse totale du nuage soit 1.

$$w(i) = w'(i)/\sum_{i=0}^{N} w'(i)$$

L'état le plus vraisemblable du système à l'instant t peut être calculé comme l'espérance mathémathique

de la distribution $p(x|z_1, z_2...z_t)$ et être approximée par l'expression

$$\sum_{i=0}^{N} x_t(i).w_t(i)$$

Si un filtre se limite aux étapes que nous venons de décrire (propagation des particules en utilisant la dynamique et mise à jour des poids en utilisant les observations), il aboutit après quelques itérations à un phénomène de *dégénérescence* dans lequel un faible nombre de particules est doté d'un poids important. La majorité des particules a alors un poids quasi-nul. Ce phénomène aboutit à de mauvaises performances de suivi car on maintient des hypothèses peu vraisemblables sur le vecteur d'état. Pour pallier ce problème, la plupart des méthodes de filtre particulaire font appel à une étape de ré-échantillonnage.

I.3.2 Phase de ré-échantillonage

Le principe du ré-échantillonage est de dupliquer les particules de poids fort et de supprimer certaines particules de poids faible. Cet étape engendre nécessairement une augmentation de la variance de l'estimation appelée *Variance de MonteCarlo*. L'une des méthodes les plus populaires pour effectuer cette étape est de réaliser un ré-échantillonage stratifié. Cette méthode décrite dans [Kit96] est avantageuse car elle permet d'effectuer le rééchantillonage en un temps linéaire tout en minimisant la variance additionnelle. Dans ce type de ré-échantillonage, la probabilité de survie d'une particule est proportionnel à son poids.

I.4 Améliorations du filtre SIR

Il existe plusieurs axes pour améliorer les performances des filtres particulaires :

- Choisir une relation état-observation de manière à optimiser son temps de calcul
- Optimiser le parcours de l'espace d'état

I.4.1 Optimisation de la fonction d'importance

Nous abordons ici la deuxième piste d'amélioration évoquée. Alors que l'approche CONDENSATION utilise uniquement la dynamique du système, d'autres méthodes comme celle décrite dans [IB98b] incluent également, dans la fonction d'importance, la relation état-observation et une connaissance *a priori* sur la distribution du vecteur d'état. La fonction d'importance peut ainsi être écrite comme une combinaison linéaire de ces trois fonctions d'importance.

$$q(x_k|x_{k-1}, z_k) = (1 - \alpha - \beta).p(x_k|x_{k-1}) + \alpha.p(x_k|z_k) + \beta.p_0(x_k) \quad avec \quad \alpha, \beta \in \mathbb{R}$$

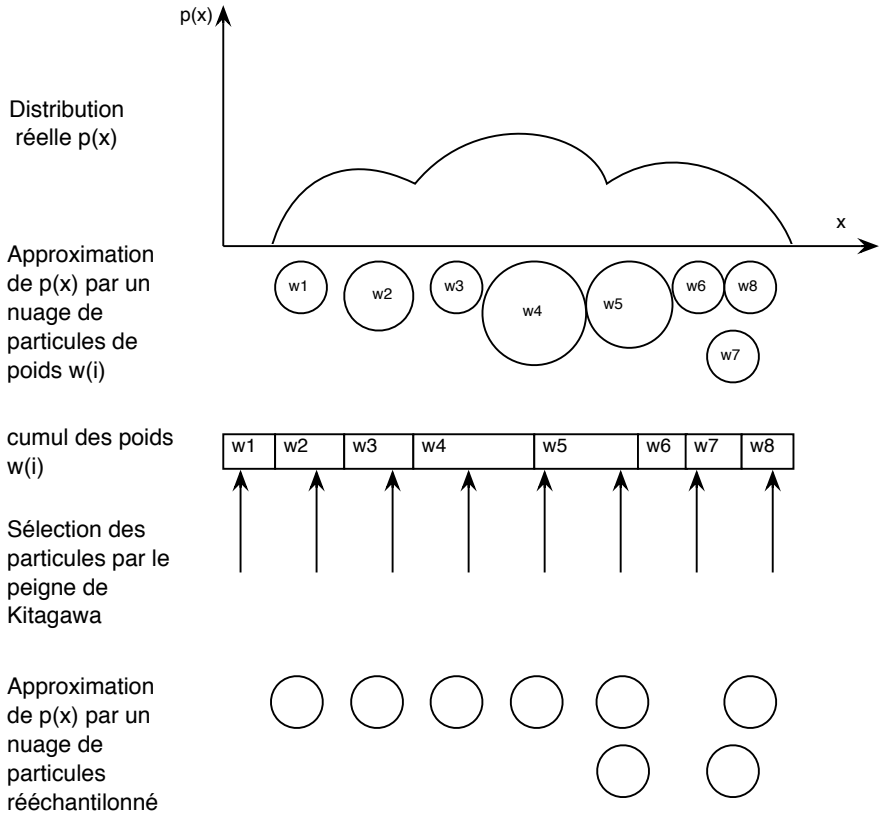


FIGURE I.2 – Illustration de l'échantillonnage proposé par Kitagawa. Dans notre exemple, la particule 5 est dupliquée et la particule 6 est supprimée.

I.4.2 Partitionnement de l'espace d'état

La mise en œuvre de la méthode SIR implique un nombre de particules qui varie exponentiellement avec le nombre de composantes du vecteur d'état pour couvrir convenablement l'espace E.

Il est possible, comme il est proposé dans [MI00], d'optimiser le parcours de l'espace d'état si la relation état-observation de la composante k du vecteur d'observation peut être réécrite ou approximée de la manière suivante :

$$p(x^k|z) = p(x^k|x^{k-1}, x^{k-2}, \ldots, x^1, z)$$

Il est alors possible de mettre à jour le nuage de particules composante par composante. Cela revient à traiter successivement la mise à jour de K filtres particulaires avec un état à une dimension[2] :

Filtre 1
Etat initial $P_0(x^1)$

[2]Un partitionnement peut également être réalisé en exprimant l'espace E comme une composition d'espaces de dimensions supérieures à 1.

Dynamique du système $p(x_t^1|x_{t-1})$
Relation état-observation $p(x^1|z)$

Filtre 2
Etat initial $P_0(x^2)$
Dynamique du système $p(x_t^2|x_{t-1})$
Relation état-observation $p(x^2|x^1, z)$

...

Filtre N
Etat initial $P_0(x^K)$
Dynamique du système $p(x_t^K|x_{t-1})$
Relation état-observation $p(x^K|x^K, x^{K-1}, ..., x^1, z)$

I.4.3 Utilisation de recuit

Une autre optimisation du filtre particulaire peut être obtenue par l'intégration de phases de recuit [DBR00] dans le but d'accélérer la convergence du nuage de particules. Nous noterons J le nombre d'itérations du filre et j l'index de l'itération. Les relations observation-état et dynamique du système sont modifiées comme suit :

1. $p(x_t|x_{t-1})(j) = p(x_t|x_{t-1})^{\alpha(j)} \qquad \alpha(j) > 1 \qquad \alpha(j+1) > \alpha(j)$
2. $p(z_t|x_t)(j) = p(z_t|x_t)^{\beta(j)} \qquad \beta(j) < 1 \qquad \beta(j+1) > \beta(j) \qquad \lim_{j \to +\infty} \beta(j) = 1$

Il est possible de proposer les interpétations suivantes de ces expressions :

1. La dynamique du système est amortie au fil des itérations.

2. Plus les itérations augmentent, plus les observations sont prises en compte (plus la sélection des particules lors des phase de rééchantillonage est importante).

Les phases de recuit permettent une meilleure convergence du filtre particulaire tout en permettant qu'il ne converge pas vers des minima locaux.

I.4.4 Optimisation de la répartition des particules par génération de nombres pseudo-aléatoires

Une autre amélioration des filtres particulaires consiste à optimiser la répartition des particules pour couvrir au mieux l'espace d'état E. Le danger d'une génération aléatoire de particules est de créer de *amas* et des *trous* (fig. I.1 b et c). Pour éviter cet écueil, il est possible de générer les nouvelles particules de manière pseudo aléatoire de manière à tendre vers une répartition plus uniforme des particules (fig. I.1 d) comme dans [GW06]. Les méthodes de cette famille sont appelées méthodes de *Quasi Monte Carlo*.

I.4.5 Présentation des différentes méthodes de filtrages

La plupart des méthodes peuvent être décrites comme l'amélioration du filtre SIR à l'aide des optimisations que nous venons de décrire. Nous synthétisons ces méthodes dans le tableau I.3.

Nom du filtre	IMP	PART	REC	QRS
SIR [Gor93]				
ICONDENSATION [IB98b]	x			
Filtre partitionné [MI00]		x		
APF [DBR00]			x	
IAPF [FLD08]	x		x	
Fitre QRS [GW06]				x
Fitre partitionné QRS [Fon08]		x		x

FIGURE I.3 – Comparaison des différents filtres particulaires présents dans la littérature. Les différentes améliorations opérées sur les filtres sont : l'amélioration de la fonction d'importance (IMP), le partitionnement de l'espace d'états E (PART), la présence de phases de recuit (REC), l'échantillonage pseudo-aléatoire (QRS)

ANNEXE II

IMAGES INTÉGRALES

Les images intégrales ne sont pas à proprement parler une méthode de traitement d'image, mais plus exactement une méthode pour optimiser le temps de calcul d'intégrales de valeurs de pixels. Les lecteurs souhaitant en savoir plus sur les traitements à base d'images intégrales peuvent se reporter à [Der07].

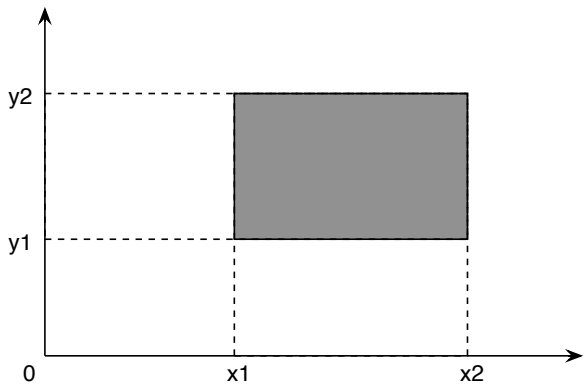


FIGURE II.1 – Schéma de l'espace d'intégration dans l'image A

Soit une image A contenant uniquement des niveaux de gris. On note $p_a(x, y)$ la valeur contenue dans le pixel de coordonnée (x, y). L'image intégrale B contient dans son pixel $p_b(x, y)$, la somme des valeurs des pixels de l'image A compris dans le rectangle ayant comme sommets l'origine de l'image $(0, 0)$ ainsi que le point (x, y). Ceci peut être résumé dans la formule suivante :

$$p_b(x, y) = \sum_{i=0}^{x} \sum_{j=0}^{y} p_a(i, j)$$

Le calcul de l'image intégrale prend un temps proportionnel à la taille de l'image si il est effectué par récurrence de la manière suivante :

$p_b(i, 0) = p_b(j, 0) = 0$
$p_b(i + 1, j) = p_b(i, j) + p_a(i + 1, j)$
$p_b(i, j + 1) = p_b(i, j) + p_a(i, j + 1)$

Une fois que l'image intégrale est calculée, le calcul d'une intégrale dans un rectangle orienté parallèlement aux axes des abscisses et des ordonnées de l'image A devient beaucoup plus simple. Il

peut être déterminé par l'intermédiaire de la formule suivante :

$$\sum_{i=x_1}^{x_2}\sum_{j=y_1}^{y_2} p_a(i,j) =$$

$$\sum_{i=0}^{x_2}\sum_{j=0}^{y_2} p_a(i,j) + \sum_{i=0}^{x_1-1}\sum_{j=0}^{y_1-1} p_a(i,j) - \sum_{i=0}^{x_2}\sum_{j=0}^{y_1-1} p_a(i,j) - \sum_{i=0}^{x_1-1}\sum_{j=0}^{y_2} p_a(i,j)$$

Soit en utilisant les images intégrales :

$$\sum_{i=x_1}^{x_2}\sum_{j=y_1}^{y_2} p_a(i,j) = p_b(x_2,y_2) + p_b(x_1-1,y_1-1) - p_b(x_2,y_1-1) - p_b(x_2,y_1-1)$$

Un calcul d'intégral dans un rectangle de l'image A orienté selon les axes d'abscisse et d'ordonnée revient donc à effectuer une somme de quatre valeurs contenues dans les pixels de l'image B.

ANNEXE III

MÉTHODES DE RECONNAISSANCE ET DE RECALAGE TEMPOREL BASÉES SUR LA PROGRAMMATION DYNAMIQUE

III.1 Dynamic Time Warping (Déformation Temporelle Dynamique)

Les méthodes de Déformation Temporelle Dynamique permettent d'obtenir simultanément un alignement temporel d'une séquence temporelle Y en de N états $[y_0, ..., y_j, ..., y_N]$ sur une séquence temporelle X de M états $[x_0, ..., x_i, ..., x_M]$, et de déterminer une distance $D(X, Y)$ entre ces deux séquences. La longueur des deux séquences peut être différente et la déformation temporelle doit préserver les relations d'antériorités entre les états de chaque séquence. Les états peuvent être exprimés dans un espace à plusieurs dimensions et sont dans un domaine continu.
Il est nécessaire de définir une métrique permettant d'estimer une distance entre un état x_i et un état y_i.

De manière imagée, le problème de détermination d'une déformation temporelle dynamique peut être formulé comme un problème de recherche de chemin optimal. Dans l'illustration III.1, sont représentées les distances entre les états x_i et y_i. Les seuls déplacements autorisés dans le tableau sont l'incrémentation de l'abscisse et/ou de l'ordonnée de 1. Nous cherchons le chemin qui minimise la somme des $d(x, y)$ des cases traversée et va de la case (x_0, y_0) à la case (x_M, y_N).

La résolution de ce problème en calculant séparément les distances cumulées pour tous les chemins possibles est extrêmement coûteuse en temps de calcul. Les calculs peuvent être optimisés en utilisant la programmation dynamique. Nous noterons $D(X_i, Y_j)$ la distance entre les séquences d'états $[x_0...x_i]$ et $[y_0...y_j]$ correspondant à une déformation temporelle dynamique optimale.

Y_6	4	1	7	0	3	1
Y_5	2	5	6	11	12	8
Y_4	6	8	3	6	9	8
Y_3	3	1	7	10	13	6
Y_2	7	0	5	3	4	0
Y_1	9	2	3	2	5	1
Y_0	1	6	9	8	6	7
	X_0	X_1	X_2	X_3	X_4	X_5

FIGURE III.1 – Distances entre les états x_i et y_j

Initialisation

$$D(X_0, Y_0) = d(x_0, y_0)$$

Récurrence

$$D(X_i, Y_j) = min(D(X_{i-1}, Y_j), D(X_{i-1}, Y_{j-1}), D(X_i, Y_{j-1})) + d(x_i, y_j)$$

Les coordonnées (x, y) du meilleur antécédent de l'état (x_i, y_j) sont notées $a(x_i, y_j)$. Les meilleurs antécédents sont représentés par des vecteurs sur la figure III.2.

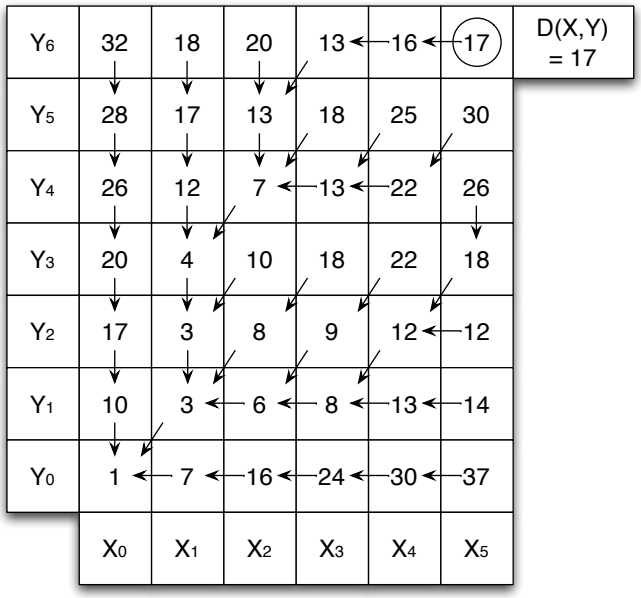


FIGURE III.2 – Distances correspondant aux déformations temporelles dynamiques et meilleurs antécédants

Détermination de la meilleure déformation temporelle

Il suffit finalement, pour déterminer la meilleure déformation temporelle de partir de l'association d'état (x_M, y_N) et de suivre les meilleurs antécédents pour trouver la meilleure déformation temporelle (fig. III.3).

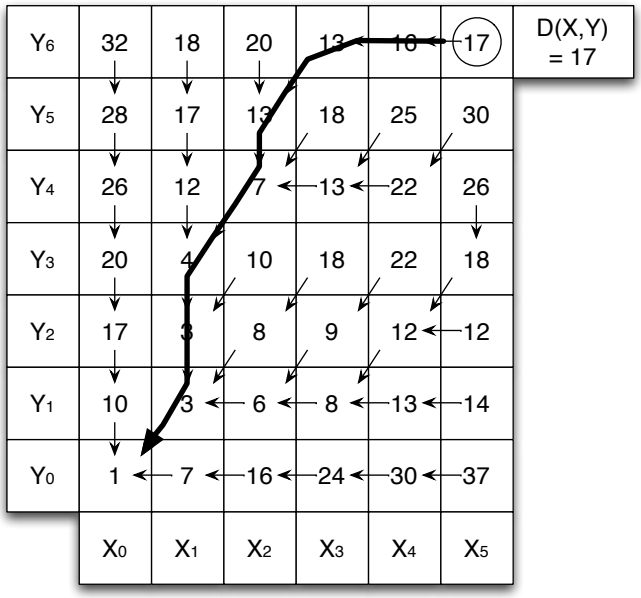


FIGURE III.3 – Meilleur chemin obtenu en suivant les meilleurs antécédants

III.2 Modèles de Markov Cachés

Nous souhaitons ici décrire brièvement le principe des Modèles de Markov Cachés (MMC) auxquels nous faisons fréquemment référence dans notre réflexion sur les différents modèles utilisés pour la reconnaissance de signes. Les explications qui suivent sont basées sur le tutoriel de Rabiner [Rab89] et son résumé en français par [LR03].

III.2.1 Fonctionnement du modèle

Un Modèle de Markov est une modélisation approchée d'un processus par un automate d'états finis à M états. L'état suivant ne dépend que de l'état précédent pour des MMC d'ordre 1. Toutefois, des modifications peuvent être apportées au modèle pour prendre en compte les N états précédents dans des MMC d'ordre supérieur.
Nous noterons s_t l'état du système à l'instant t et $a(m, m')$ la probabilité que l'automate se retrouve à l'état m' à l'instant t s'il se trouvait à l'état m à l'instant $t-1$. A l'instant initial, la distribution connue de l'état de système est notée $d_0(m)$.
Lorsque l'automate passe dans l'état m, il génère une valeur y. Les valeurs de sorties sont discrètes et sont numérotées de 1 à N.
La probabilité d'émission de la $n^{ième}$ valeur n quand le système se trouve dans l'état m est noté $b(m, n)$.

Un modèle de Markov sera donc défini par les paramètres suivants :

- La matrice A de transition d'état du système,
- La matrice B de génération de valeur à partir d'un état,
- La distribution initiale D_0 des états du système.

Le modèle de Markov est dit "caché" dans la mesure où on ne connaît avec certitude que les observations du système qui ne reflètent qu'indirectement les états du système. Par ailleurs, les valeurs générées à partir d'un état sont émises de manière probabiliste.
Il est important de noter que l'état du système à l'instant t ne dépend que de son état à l'instant $t-1$. Il n'y a ainsi pas de mémoire de l'ensemble des états antérieurs

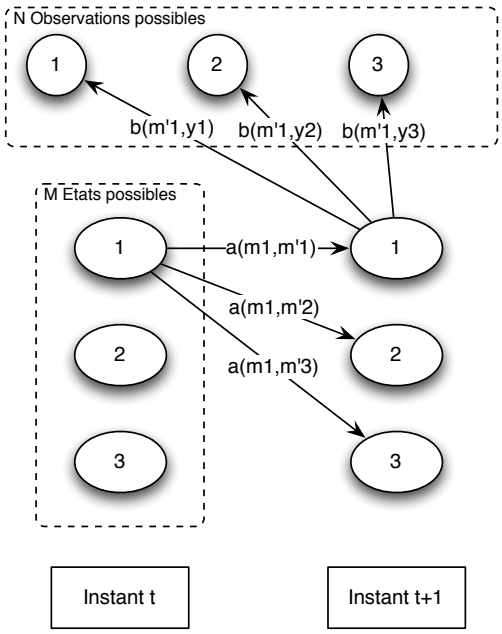


FIGURE III.4 – Schéma récapitulatif des Modèles de Markov Cachés

III.2.2 Utilisation des modèles de Markov pour la reconnaissance

Dans le cas de l'utilisation d'un MMC pour la reconnaissance, on cherche à trouver la séquence d'états la plus probable $F = [f_0, ..., f_t, ..., f_T]$ qui aurait pu générer une séquence d'observation qu'on connaît $Y = [y_0, ..., y_t, ..., y_T]$.
Il est possible de déterminer par récurrence la probabilité r_T d'obtenir la séquence Y à l'aide de notre modèle. Lorsque les MMC sont utilisées pour effectuer une reconnaissance, le MCC ayant la probabilité la plus élevée est considéré comme étant le meilleur modèle expliquant les observations. Nous noterons $r_t(m)$ la probabilité de génération de la séquence d'observations $Y_t = [y_0, y_1, ..., y_t]$ avec la séquence d'état la plus probable se finissant par l'état m.

Initialisation (A l'instant $t_0 = 0$))

$$r_0(m) = d(m).b(m, y_0)$$

Récurrence

$$r_t(m') = \max_m r_{t-1}(m)a(m, m')b(m', y_t)$$

Il est ainsi possible de noter le meilleur prédécesseur de m' à l'instant t que nous nommerons $q_t(m')$.

Choix du dernier état

L'état f_T retenu à l'instant T est celui qui maximise $r_T(m)$.

Détermination de la séquence la plus probable
On procède également par récurrence en utilisant les meilleurs prédécesseurs f_{t-1}.

$$f_{t-1} = q_t(f_t)$$

III.2.3 Détermination de la probabilité d'observation d'une séquence

Dans la section précédente, r_T représentait la probabilité de générer la séquence d'observation Y avec la séquence d'observation la plus probable. Nous souhaitons maintenant calculer la probabilité d'obtenir Y avec notre modèle, en prenant en compte toutes les séquences d'état possible.

On nomme $P(Y/S)$ la probabilité de générer la séquence d'observations Y à partir de la séquence d'états S.

$$P(Y/S) = \prod_{t=0}^{T} b(s_t, y_t)$$

On nomme $P(S)$ la probabilité de générer la séquence d'état $S = [s_0, s_1, ..., s_T]$ à partir du modèle de Markov.

$$P(S) = d_0(s_0) \prod_{t=1}^{T} a(s_{t-1}, s_t)$$

On nomme $p(Y, S)$ la probabilité de générer la séquence d'état S et la séquence d'observation Y.

$$P(Y, S) = P(Y/S).P(S) = d_0(s_0).b(s_0, y_0) \prod_{t=1}^{T} a(s_{t-1}, s_t).b(s_t, y_t)$$

Il suffit ensuite de sommer les probabilités d'obtenir l'observation Y avec chaque séquence d'état S_i possible.

$$P(Y) = \sum_{S_i} P(Y, S_i)$$

Le problème est qu'il existe M^T différentes séquences d'états possibles. Il en résulte une quantité de calculs prohibitive. $P(Y)$ peut heureusement être calculé par récurrence.
Dans la suite des calculs, nous noterons $P(Y_t, m)$ la probabilité d'obtenir la séquence d'observation partielle $Y = [y_0...y_t]$ avec n'importe quelle séquence d'état se finissant par l'état m.

Initialisation

$$P(Y_0, m) = d(m).b(m, y_0).$$

Récurrence

$$P(Y_{t+1}, m') = \sum_{m \in M} P(Y_t, m).a(m, m').b(m', y_{t+1})$$

Fin

$$P(Y_T) = \sum_{m \in M} P(Y_T, m)$$

Il est également possible de calculer la probabilité qu'a le système de se trouver à l'état q à un instant t quelconque par programmation dynamique. Nous ne faisons pas référence à cet algorithme dans le manuscrit et nous n'avons pas utilisé cet algorithme dans notre travail. Les lecteurs souhaitant connaître l'algorithme sont invités à se reporter à [Rab89].

III.2.4 Estimation des paramètres optimaux du Modèle de Markov Caché

Le but est maintenant d'estimer les paramètres optimaux du Modèle de Markov. Ce modèle est constitué à partir d'un ensemble de J observations Y_j. L'objectif de l'algorithme est d'optimiser la vraisemblance des séquences d'après le Modèle de Markov.
L'algorithme de Baum Welch qui réalise cette tâche est décrit dans [BPSW70]. Il procède de manière itérative en alternant deux étapes :

- L'évaluation de probabilité de passage dans les différents états m à chaque instant t à partir des observations Y_j et du modèle défini par A, B et D_0.
- La réévaluation du modèle (A, B et D_0) à partir de la succession de distribution d'états déterminée à l'étape précédente.

III.3 Comparaison des méthodes DTW et HMM

Nous venons de décrire deux familles de méthodes basées sur la programmation dynamique et permettant d'effectuer simultanément un recalage temporel et une reconnaissance de séquences temporelles. Le tableau qui suit résume les avantages et les inconvénients de ces deux approches.

Méthode	Avantages	Inconvénients
DTW	Peu d'apprentissage nécessaire Différentes normes utilisables	Modèle volumineux Temps de calcul important
MMC	Modèle léger Topologie adaptable Rapide à calculer	Grand corpus d'apprentissage nécessaire

FIGURE III.5 – Tableau comparatif des Modèles de Markov Cachés et des Déformations Temporelles Dynamiques